

**QUANTIFYING HUMAN PERFORMANCE OF A DYNAMIC
MILITARY TARGET DETECTION TASK: AN
APPLICATION OF THE THEORY OF
SIGNAL DETECTION**

STUART L. TURNER

**CREW SYSTEMS DIRECTORATE
HUMAN ENGINEERING DIVISION
WRIGHT-PATTERSON AFB, OHIO 45433-7022**

INTERIM REPORT FOR THE PERIOD NOVEMBER 1994 TO JUNE 1995

June 1995

Approved for public release; distribution is unlimited

**AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573**

**ARMSTRONG
LABORATORY**

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Armstrong Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

TECHNICAL REVIEW AND APPROVAL


AL/CF-TR-1995-0130

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

The voluntary informed consent of the subjects used in this research was obtained as required by Air Force Instruction 40-402.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER


KENNETH R. BOFF, Chief
Human Engineering Division
Armstrong Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 1995		3. REPORT TYPE AND DATES COVERED Interim Report Nov 94 - Jun 95	
4. TITLE AND SUBTITLE Quantifying Human Performance of a Dynamic Military Target Detection Task: An Application of the Theory of Signal Detection				5. FUNDING NUMBERS PE: 62202F PR: 7184 TA: 10 WU: 45	
6. AUTHOR(S) Stuart L. Turner					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Armstrong Laboratory, Crew Systems Directorate Human Engineering Division Human Systems Center Air Force Materiel Command Wright-Patterson AFB OH 45433-7022				10. SPONSORING/MONITORING AGENCY REPORT NUMBER AL/CF-TR-1995-0130	
11. SUPPLEMENTARY NOTES					
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) As crew aiding technologies are developed to assist military aviators in performing complex target detection tasks, evaluation metrics must be developed which are common to both human operators and automatic target recognition (ATR) systems so that performance comparisons can be efficiently conducted. The dynamic nature of the multiple target detection task introduces several unique problems in quantifying detection performance. Classical methods of implementing the Theory of Signal Detection (TSD) to quantify performance have proven to be insufficient, and ATR evaluators have developed unique metrics which have not been applicable to evaluating human performance. This research introduced a novel application of TSD to the dynamic, multiple target detection scenario, and a new method of evaluating human performance was developed by adapting an established ATR evaluation method to human subject performance. A linear relationship was discovered between the TSD metric d'e and the new evaluation metric, validating the new evaluation method. The new method provided a common metric for evaluating both human and ATR performance of multiple target detection tasks.					
14. SUBJECT TERMS Target Detection Automatic Target Recognizer LANTIRN Signal Detection Infrared FRACTIL ATR				15. NUMBER OF PAGES 102	
				16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNLIMITED		

THIS PAGE INTENTIONALLY LEFT BLANK

PREFACE

This work was performed under Air Force Work Unit 71841045, "Strategic Crew Performance." It was performed by the Crew Systems Integration Branch, Human Engineering Division, Crew Systems Directorate, of Armstrong Laboratory.

This work was a follow-on to a research effort conducted by Capt Stuart L. Turner and Mr. Bradley D. Purvis of the Human Engineering Division. The initial research involved a field evaluation of pilot workload and operational performance associated with the employment of a prototype automatic target recognition device which was demonstrated in a LANTIRN targeting pod on an F-16 fighter in 1993. The Crew Systems Integration Branch participated in the initial technology demonstration along with Wright Laboratory's Sensor Evaluation Branch and the LANTIRN System Program Office. Data collected during the technology demonstration facilitated this follow-on research conducted at the Human Engineering Division.

Special recognition is accorded the following individuals who consulted on or assisted with this project:

Mr. Bradley D. Purvis, Lt Col William Wittman, Mr. Gilbert Kuperman, and Ms. Denise Wilson of the Armstrong Laboratory, Human Engineering Division, Wright-Patterson Air Force Base, Ohio.

Mr. Lloyd Clark and Ms. Lori Westercamp of the Wright Laboratory, Wright-Patterson Air Force Base, Ohio.

Ms. Iris Davis, Mr. Chuck Goodyear, and Ms. Elisabeth Fitzhugh of Logicon Technical Services, Dayton, Ohio.

Mr. Greg Bothe, Sr., Mr. Thomas Haberlandt, and Mr. Steve Ellis of Science Applications International Corporation, Dayton, Ohio.

Dr. Jennie J. Gallimore, Dr. Anthony J. Cacioppo, and Dr. Daniel L. Weber of Wright State University, Dayton, Ohio.

SUMMARY

AL/CFHI participated with Wright Laboratory and the LANTIRN System Program Office in a technology demonstration of automatic target recognition systems in 1993. After a field study of the impact of the automatic target recognizer (ATR) on combat performance and pilot workload, Wright Laboratory assumed responsibility for conducting laboratory tests of additional ATR devices. Using infrared imagery of low level ingress passes on military ground targets collected during the technology demonstration, ATR evaluations were begun in the laboratory. These evaluations would compare the target detection performance among the ATR devices, but no comparison was planned between ATR performance and human detection performance. The Crew Systems Integration Branch assumed the task of evaluating human detection capability in a manner that would allow direct comparison with ATR performance results.

Because the ATR evaluations had already begun and a performance metric had already been selected by the ATR evaluators, a means of evaluating human subjects using the same ATR performance metric was generated. This new human evaluation method would allow human-ATR comparisons using common stimuli and a common performance metric. However, the new human evaluation method was unproved and needed to be validated. The well established Theory of Signal Detection (TSD) was employed simultaneously with the new evaluation method as a validation method.

Target detection performance of twelve human subjects viewing the technology demonstration IR imagery was quantified using both TSD techniques and the new evaluation technique. The experiment was conducted in the laboratory using high-fidelity video displays, computer generated graphics, and computer controlled presentations. The dependent measures from each of the two evaluation techniques were subjected to a repeated measures ANOVA and Pearson Product correlation.

The results quantified human target detection performance using an infrared sensor as a function of range to target and combined atmospheric and ground clutter conditions. Detection performance was shown to deteriorate with range to target, and performance was shown to deteriorate more severely under conditions of high atmospheric humidity. Detection performance was shown to be superior at close ranges under low ground clutter conditions. Most significantly, a strong linear relationship was found between the TSD metric and the metric derived from the new evaluation method. This relationship validates the new evaluation method and facilitates follow on work comparing the results of this research with that of the ATR evaluations.

The new evaluation method developed in this research may, with additional research, begin the development of a bridge between human and machine evaluation methods. Comparing human and machine performance will become more common as electronic crew aiding technologies are developed to assume or assist with tasks previously performed by crew members alone. New and innovative techniques will be required to fulfill the comparison requirements. This research represents one small step in that direction.

TABLE OF CONTENTS

	Page
INTRODUCTION	1
BACKGROUND	3
Operational Relevance	3
Theory of Signal Detection	7
The Receiver-Operating Characteristic Curve	14
Empirical ROC Curves and d' Metrics	19
Hit-False Alarm Count Curves	24
RELATED RESEARCH	30
LANTIRN FRACTIL Research	30
Wright Laboratory ATR Evaluations	31
Barnes	32
Ozkaptan	34
Astley, Taylor, Boggis, Asbury, and Wilson	35
SUMMARY	38
OBJECTIVES	40
Assumptions	40
Hypotheses	42
METHOD	43
Experimental Design	43
Subjects	45
Apparatus	45
Stimuli	48
Procedure	51
Data Reduction	57
Hit-FAC Data Reduction	57
TSD Data Reduction	59

TABLE OF CONTENTS continued

RESULTS	63
Data Analysis	63
TSD Technique (d'_e variable)	63
Hit-FAC Technique (HR variable)	66
Correlations	70
DISCUSSION	77
Hypotheses	77
Summary	80
CONCLUSIONS	84
APPENDIX	86
REFERENCES	91

LIST OF FIGURES

Figure	Page
1. Illustration of operational scenario range bins	6
2. Theoretical frequency distribution of noise, $f_n(x)$, and signal + noise, $f_{sn}(x)$, in the decision space	9
3. Theoretical distributions for three different values of signal strength . . .	10
4. Location of observer's response criterion (k) divides the sensory magnitude axis	12
5. The four event regions of the decision space distributions	13
6. Relationship between decision space distributions and ROC curve . . .	15
7. Relation between $f_{sn}(x)/f_n(x)$ ratio and the ROC curve	17
8. Family of ROC curves of varying d'	18
9. Relationship of subject confidence reports to decision space criterion line.	21
10. Typical hit-FAC curve for ATR performance	26
11. Example of hit-FAC curve creation from empirical data	28
12. 4 x 2 factorial repeated measures design	44
13. Block diagram of experimental apparatus	46
14. Illustration of typical display and menu for TSD stimuli and touch response	54
15. Stimulus presentation chronology	56
16. Example of hit-FAC curve creation from human subject data	60
17. d'_e vs RANGE BIN for both levels of ATMOSPHERE-CLUTTER . . .	64
18. HIT RATE vs RANGE BIN for both levels of ATMOSPHERE- CLUTTER	68

LIST OF FIGURES continued

19.	Linear regression of d'_e on HIT RATE for all points	72
20.	Linear regression of d'_e on HIT RATE for Edwards	73
21.	Linear regression of d'_e on HIT RATE for Eglin	73
22.	Linear regression of d'_e on HIT RATE for all points except Edwards BIN = 4	74
23.	Linear regression of d'_e on RANGE BIN for each ATMOSPHERE- CLUTTER	75
24.	Linear regression of HIT RATE on RANGE BIN for each ATMOSPHERE-CLUTTER	75
25.	d'_e as a function of RANGE BIN and ATMOSPHERE-CLUTTER with standard error of mean	82
26.	HIT RATE as a function of RANGE BIN and ATMOSPHERE- CLUTTER with standard error of mean.	82
27.	Sample ROC curves for Edwards RANGE BIN 2	87
28.	Sample ROC curves for Edwards RANGE BIN 4	87
29.	Sample ROC curves for Edwards RANGE BIN 6	88
30.	Sample ROC curves for Edwards RANGE BIN 2	88
31.	Sample hit-FAC curve for Edwards RANGE BIN 2	89
32.	Sample hit-FAC curve for Edwards RANGE BIN 4	89
33.	Sample hit-FAC curve for Edwards RANGE BIN 6	90
34.	Sample hit-FAC curve for Edwards RANGE BIN 2	90

LIST OF TABLES

Table		Page
1.	General stimulus-response matrix of estimates of conditional probability for four rating responses	22
2.	Sample stimulus-response frequency matrix for 200 trials	23
3.	Sample stimulus-response conditional probability estimates matrix for 200 trials	23
4.	Cumulative conditional probability estimates matrix for four response ratings	24
5.	Cumulative proportions of each response rating transformed into z-scores for the calculation of d'	61
6.	ANOVA summary for d'_e	65
7.	Results of simple effects analysis of RANGE BIN by ATMOSPHERE-CLUTTER using the d'_e dependent variable	66
8.	Results of simple effect analysis of ATMOSPHERE-CLUTTER by RANGE BIN using the d'_e dependent variable	67
9.	ANOVA summary for HR	69
10.	Results of simple effects analysis of RANGE BIN by ATMOSPHERE-CLUTTER using HR dependent variable	69
11.	Results of simple effects analysis of ATMOSPHERE-CLUTTER by RANGE BIN using the HR dependent variable.	70

INTRODUCTION

Modern electronics technology has spawned a new class of research and development known as "crew aiding". Crew aiding technologies are defined loosely as technologies which aid or assist a system operator in making intelligent, informed decisions about the operation of the system, which enhance the mental or physical capabilities of the operator, or which automatically perform tasks for which the higher cognitive capability of the human operator is not required. Human factors researchers are interested in maximizing the performance enhancements that crew aiding offers by refining the operator interfaces to ensure safe and efficient utilization of crew aiding technology.

Crew aiding technologies are rapidly being applied by the military in the operation of a wide variety of devices, particularly in combat aircraft. One crew aiding technology currently under development is the automatic target recognizer (ATR). The ATR is envisioned as an aid to the pilot in detecting and recognizing targets well beyond the visual capabilities of the pilot. The pilot-ATR system performance of target detection tasks is expected to be superior to the performance of the unaided pilot. In order to effectively quantify the impact of ATR technology on human-system performance of a target detection task, a baseline of unaided human performance of the task must first be determined.

While several methods could be employed to describe human performance of a detection task, the Theory of Signal Detection (TSD) is most widely used in modern research. TSD facilitates the generation of receiver operating characteristic (ROC) curves which reflect the probability of detecting a target (hit) and the probability of incorrectly reporting a target (false alarm) along a continuum of receiver response bias. TSD has been successfully used to describe the human receiver under limited, well controlled, laboratory conditions. Laboratory conditions typically involve static image stimuli containing either a single target

opportunity or no target opportunity. Subjects evaluate numerous stimuli under variable biasing rules in order for an ROC curve to be constructed.

However, current TSD techniques are not well suited for the evaluation of the human receiver under more dynamic and uncontrolled conditions such as those associated with the pilot's scenario. The pilot must evaluate a dynamic visual display of terrain scenery which may contain multiple targets within any given scene. The pilot is time-limited in his decision process since the opportunity to strike a target may be a fleeting one. New and innovative TSD techniques are required to assess performance under these conditions.

Military ATR evaluators have developed metrics that facilitate the use of dynamic video imagery as the input stimulus to ATR systems. Performance curves are plotted which relate target hit rates to absolute false alarm counts, (FAC) rather than false alarm probabilities. This type of performance curve (hit-FAC curve) is one of the accepted standard metrics for ATR evaluation by the U.S. military. The hit-FAC curve is similar to the ROC curve in its shape and graphical interpretation, but it is derived through techniques different from those of TSD. A comparison of ATR performance with human performance can best be made using this type of curve as a common means of describing performance.

The current project is concerned with baselining human performance of an operationally derived target detection task using an innovative technique for generating TSD metrics with dynamic video stimuli. Further, a new technique for generating hit-FAC curves for human subjects will be validated by correlating TSD metrics with metrics derived from the human hit-FAC curves. Ultimately, the validated hit-FAC curves will facilitate a follow-on comparison of human performance with the performance of three different ATR designs which have been previously evaluated using the hit-FAC curve technique.

The initial topic is a description of the operational scenario from which the detection task is derived. This is followed by an overview of signal detection theory, a review of related research literature, and the method employed in this research.

BACKGROUND

Operational Relevance

War-fighting experience in Operation Desert Storm proved that locating and striking mobile ground targets from the air is a very difficult task, even in open, barren desert terrain. Mobile ground targets are mobile missile launchers, surface-to-air threats, tanks, and other support vehicles which can relocate autonomously. Developing technologies which improve the ability to place mobile targets at risk has become a primary defense department goal.

Mobile target detection technologies involve both active airborne sensors, such as radar, and passive airborne sensors, such as Forward Looking Infrared (FLIR) or common optical imaging techniques. Current operational concepts call for these sensors to be employed on high flying surveillance platforms which perform initial detection of targets and pass navigational information to a strike aircraft, typically a fighter-bomber. The strike aircraft directs itself to the designated coordinates to destroy the targets. However, because the targets are mobile and because the target position information may be somewhat inaccurate or old, the strike aircraft must perform its own search and detection task within an area of uncertainty determined by the quality of the target data.

The strike aircraft's search task may start with an active radar search, but it must ultimately employ some infrared or optical imaging technique to confirm the target location and identity. This "end game" search task is often difficult due to the nature of the sensors employed. Substantial image magnification is required to maintain a safe distance from the targets while launching munitions against them. One commonly used targeting system, the Low Altitude Navigation and Targeting Infrared for Night (LANTIRN), provides a targeting magnification that results in a search cone just 1.67 degrees wide. The aircrews describe the employment of this sensor as "looking through a soda straw." Even at extreme magnification with

optics on-target, detecting and recognizing targets at extended range is a difficult task in a visually cluttered, dynamic terrain scene.

This end-game search task has been identified by the U.S. Air Force as a candidate for crew aiding technology to assist in the detection and recognition of targets. Automatic target recognizers are being developed to analyze imagery, detect targets, and provide an assessment of the target type to the aircrew. Prototype ATR's have been operationally evaluated in Air Force advanced technology demonstrations. One such demonstration conducted in 1993 inserted an ATR device into a LANTIRN system employed on an F-16 fighter aircraft. The effort, known as FLIR And Crier Technology Insertion into LANTIRN (FRACTIL), sought simply to prove the operational concept for employing ATR technology in a low-level flight environment.

A secondary goal of FRACTIL was to collect high resolution video data of ground targets which could be used in the laboratory to conduct well controlled evaluations and comparisons of ATR devices. The aircraft was modified with a digital video recorder known as the Digital Cassette Recording System with incremental tape motion (DCRSi). This system recorded the FLIR imagery from the LANTIRN pod as the aircraft flew multiple, low-level, ingress passes over variable ground arrays of tanks, military support vehicles, and other mobile targets. Several hours of high-resolution video data were recorded at two test range sites: (1) Eglin Air Force Base (AFB), Florida, which offered high atmospheric humidity, high visual clutter, and low thermal clutter; and (2) Edwards AFB, California, which offered very low atmospheric humidity, low visual clutter, and high thermal clutter.

The digital video data have been used as input stimuli by the Air Force Wright Laboratory (WL) in a performance evaluation of three different ATR designs. Plots of ATR performance were constructed in terms of correct target detection versus false detections. These metrics facilitated performance comparisons between individual ATR designs. However, in order to assess the crew aiding value of these devices, a comparison must be made directly with the

performance of the unaided human operator in the same task. In order to conduct this comparison, a common means of evaluating both the ATR and the human must be devised.

The ATR method of evaluation was based on a community standard metric which plots the probability of achieving correct target detection (hit rate) against a count of incorrect target detections (false alarm count, FAC). As the reader will find in subsequent discussion, these plots resemble, but differ from, ROC curves. In the WL ATR evaluation, a performance plot was generated for different range "bins" since performance changes as a function of range to target. A range bin was defined by the calculated slant range to target as the aircraft flew toward the target array. Generally, bins were defined in one kilometer steps. Figure 1 illustrates the scenario.

Because the ATR evaluation did not allow TSD metrics to be derived for the ATR's, and because the hit-FAC metric employed is a community standard for ATR evaluation, the best means of comparing human and ATR performance is to generate hit-FAC metrics for human performance of the task. A bin by bin performance comparison can then be made.

However, this does not alleviate the desire to generate TSD metrics for the human. The well established TSD metrics can be used to help validate a technique for producing human hit-FAC curves, and TSD results can be used for comparison in follow-on studies of human performance. Also, the descriptive results of the FRACTIL demonstration seem to indicate that human performance of the detection task decreases markedly at longer stand-off ranges where the ATR seemed to excel. Generation of TSD measures will allow between-bin comparisons of human performance to validate this hypothesis, as well as comparisons between the atmosphere-clutter conditions.

A cursory discussion of TSD, the generation of ROC curves and the d' prime (d') measure, and the relationship of these to the hit-FAC plots is necessary to realize the unique character of this research. These topics are developed in following sections.

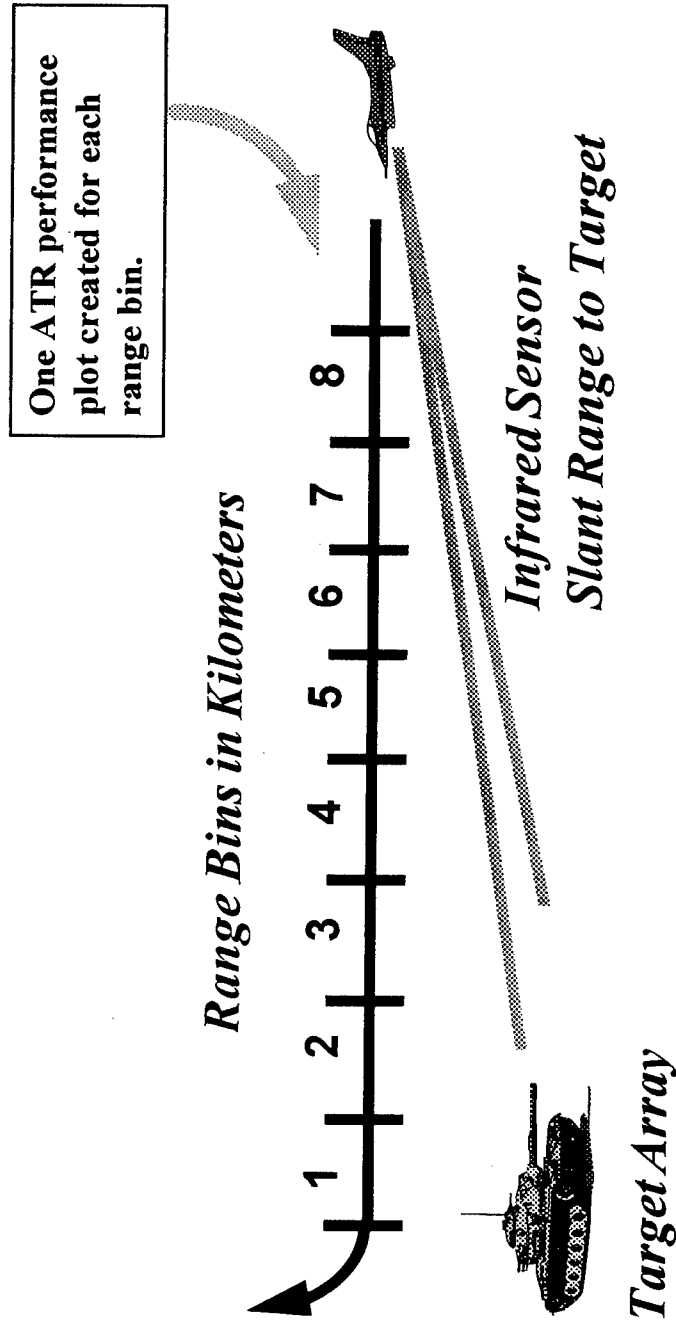


Figure 1. Illustration of operational scenario range bins.

Theory of Signal Detection

The Theory of Signal Detection was originally conceived in the early 1950's as a theoretical modeling tool for electronic signal detection. Aspects of statistical decision theory were combined with electronic signal theory to form a general theory of signal detectability (Swets, 1964). The theory was applied to psychophysical measurement as a way to explain how weak visual or auditory signals are distinguished from a cluttered or "noisy" background (Swets, 1964; Green and Swets, 1966). Experimental methods and analytical techniques were developed to separate human decision factors from sensory factors in the human's effort to optimize performance in signal detection tasks.

These efforts formed the foundation of modern detection theory's application to psychological and psychophysical research. In subsequent years, TSD has broadened to include alternative theoretical assumptions and has been applied to analyze numerous experimental tasks (Macmillan and Creelman, 1991). One of these tasks, target detection, is the subject research.

In applications of TSD, an observer always detects a signal against some background level of activity, referred to as noise. Noise is assumed to vary in a random manner. The observer is presented with a stimulus that may contain signal and noise together, or may be simply noise alone. Upon his observation (x), the observer must determine whether x is due to noise only or due to the inclusion of signal with the noise. If a weak signal is present, its detection may be difficult and the signal overlooked as noise. If a particular sample from the noise is of a large magnitude, it may be mistaken as the presence of signal (Green and Swets, 1966).

In the operational ground-target detection task, signal plus noise is equivalent to a visual image stimulus in which a target resides within a scene of variable background terrain. Noise only is a scene containing only the variable background terrain. Occasionally, a target may not stand out significantly from the background scene and the signal may be overlooked as noise. On other occasions,

a bright terrain feature, such as a rock or tree, may be incorrectly perceived as a target and result in an incorrect detection, or "false alarm".

The decision domain underlying signal detection is depicted in Figure 2. Two probability distributions are represented. The horizontal axis reflects the magnitude of the sensory observation, while the vertical axis is the probability density associated with each value of sensory magnitude. In our operational scenario, sensory magnitude may be a function of multiple factors such as the brightness, contrast, and size of objects within the scene, as well as the observer's visual-perceptual capability, which impact his interpretation of the stimulus.

Distribution $f_n(x)$ is the distribution associated with noise-only stimuli. Distribution $f_{sn}(x)$ is the distribution associated with signal-plus-noise stimuli. Because signal is added to noise, the mean sensory magnitude for the signal-plus-noise distribution is always greater than that of the noise-only distribution. Thus, the signal-plus-noise distribution is shifted to the right of the noise-only distribution along the sensory magnitude axis. Overlap of the two distributions represents sensory magnitudes which are included in both distributions.

The separation of the two distributions along the sensory magnitude axis is partially determined by the strength of the observed signal. If the signal is very strong, greater sensory magnitudes will result on average. The signal-plus-noise distribution will shift to the right increasing the distance between the distributions. If the signal is weak, the signal-plus-noise distribution will shift to the right less dramatically, or perhaps not at all, resulting in little distance between the distributions as illustrated in Figure 3.

Since the distributions are also partially a function of the sensory capability of the observer, the distance between the means of the distributions can be employed as a metric describing this sensory capability for a given set of stimulus conditions. This measure, expressed in units of the $f_n(x)$ standard deviation, is referred to as d' (Swets, 1964).

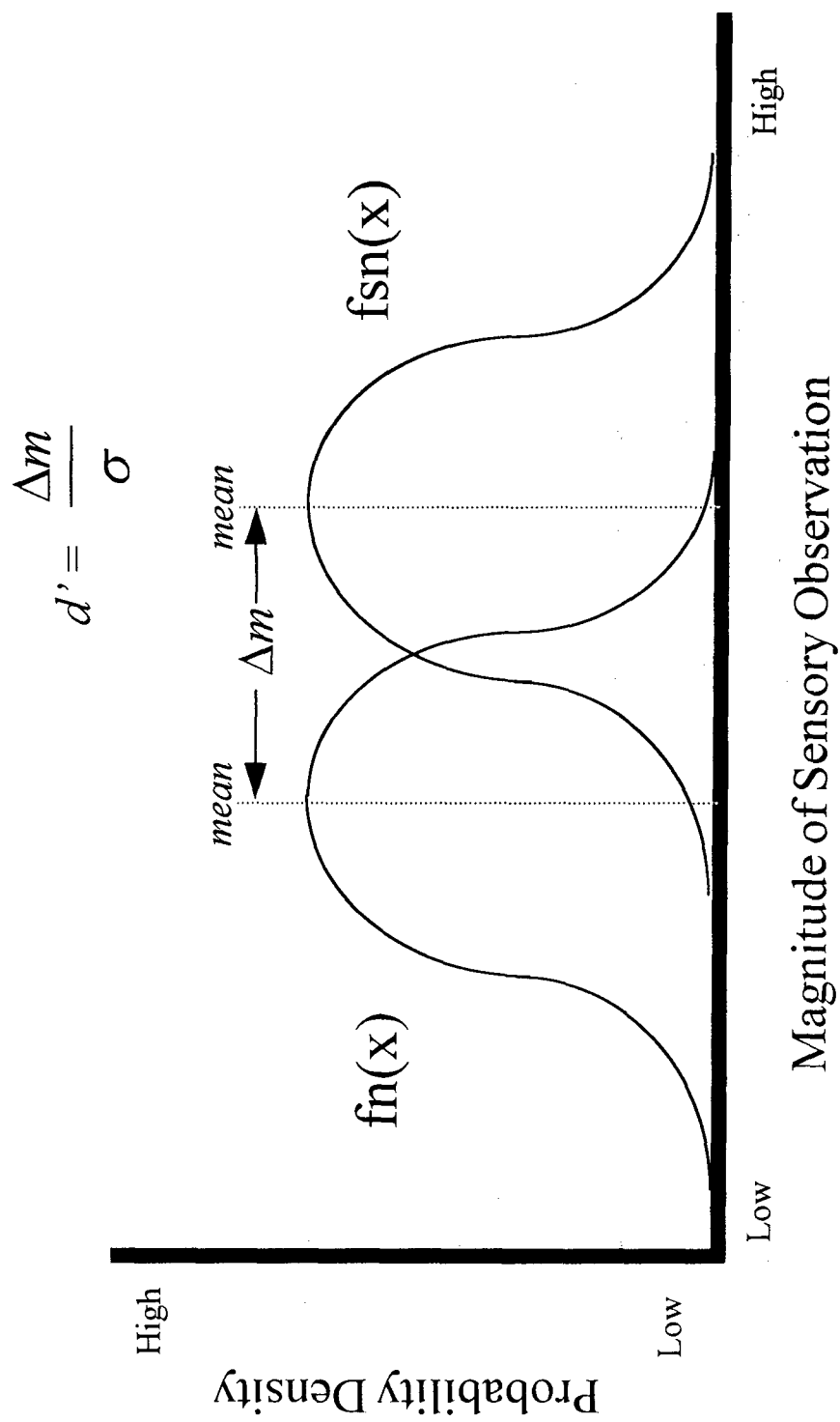


Figure 2. Theoretical frequency distribution of noise, $f_n(x)$, and signal + noise, $f_{sn}(x)$, in the decision space.

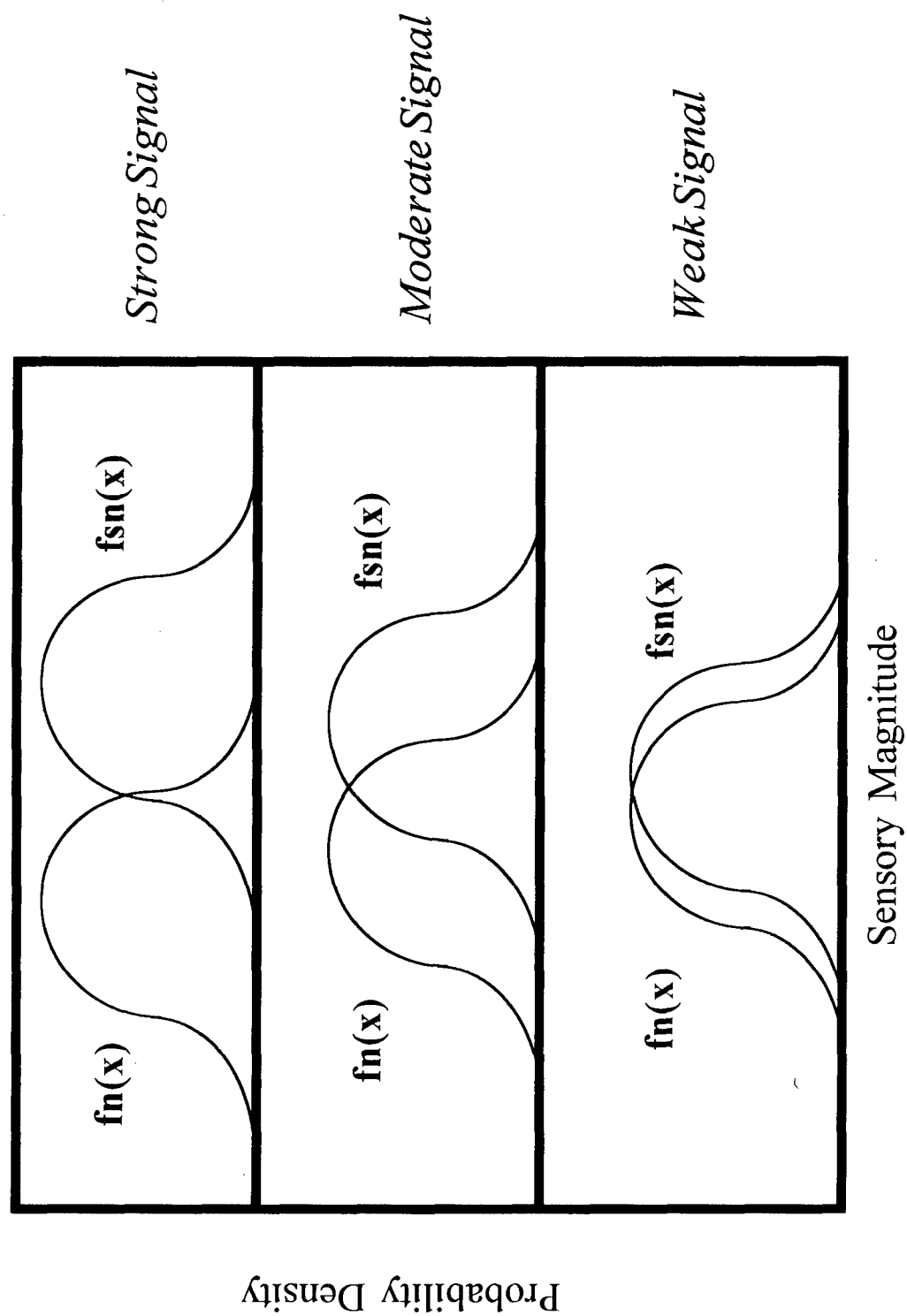


Figure 3. Theoretical frequency distributions for three different values of signal strength.

The Theory of Signal Detection assumes that the observer employs some decision-making rules, or criteria, which affect his detection performance. The observer establishes a particular rule partitioning the sensory magnitude axis into two segments as depicted in Figure 4. The value k represents the decision criterion used by the observer. All stimuli resulting in a sensory magnitude greater than k are decided to be the result of signal plus noise. All stimuli resulting in a sensory magnitude less than k are decided to be the result of noise only. The dividing value k is indicative of the observer's response *bias* (B). The response bias B is calculated as the likelihood ratio $f_{sn}(x) / f_n(x)$ when x equals the criterion value k (Swets, 1964). The reader may observe that moving the criterion value to the left decreases the value of B and moving the criterion to the right increases B .

Once again relating the operational scenario, bias may be established by the observer's rules of engagement. For example, if the pilot is provided with infinite munitions and told that his best chance for success is to fire upon anything that remotely resembles a target, his criterion value would lie far to the left on the sensory magnitude axis. A very low sensory magnitude would result in a signal-plus-noise, or "target present" decision. Conversely, if the pilot has a small and finite number of munitions and is instructed to fire upon only those objects that he is certain are hostile targets, the criterion value will lie far to the right on the sensory magnitude axis. Only very strong sensory magnitudes will produce the "target present" decision. Rules resulting in intermediate criterion values between these two extreme examples are also possible. In fact, an entire spectrum of bias is possible along the horizontal sensory magnitude axis.

With the overlapping distributions and the dividing criterion value, four stimulus-response events are possible. The four event regions of the distributions are more clearly indicated when the distributions are separated as in Figure 5. The convention of estimates of conditional probabilities is used to focus attention on the observer's behavior and to minimize the impact of the number of presentations of either stimulus condition (Green and Swets, 1966).

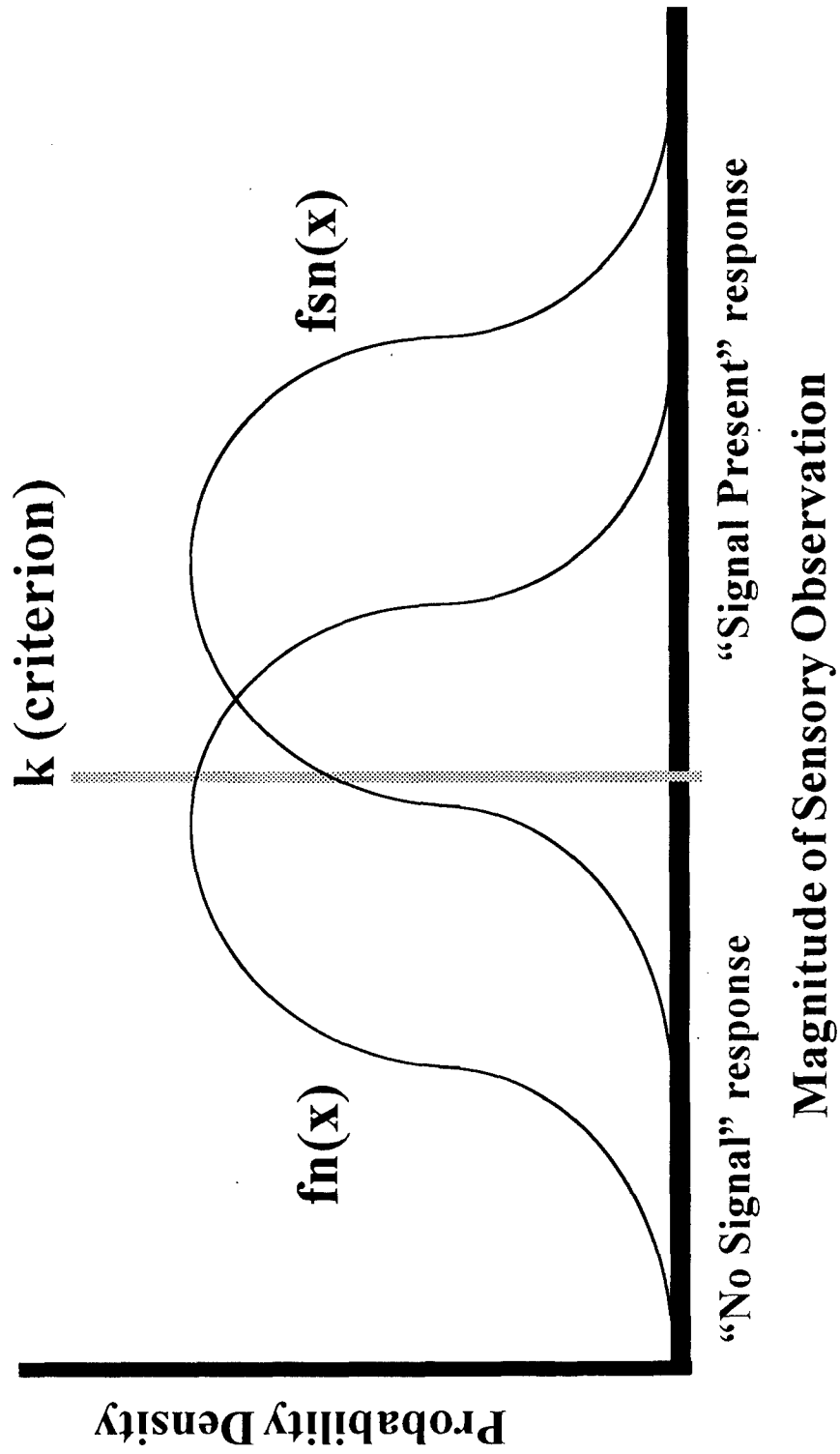


Figure 4. Location of observer's criterion (k) divides the sensory magnitude axis.

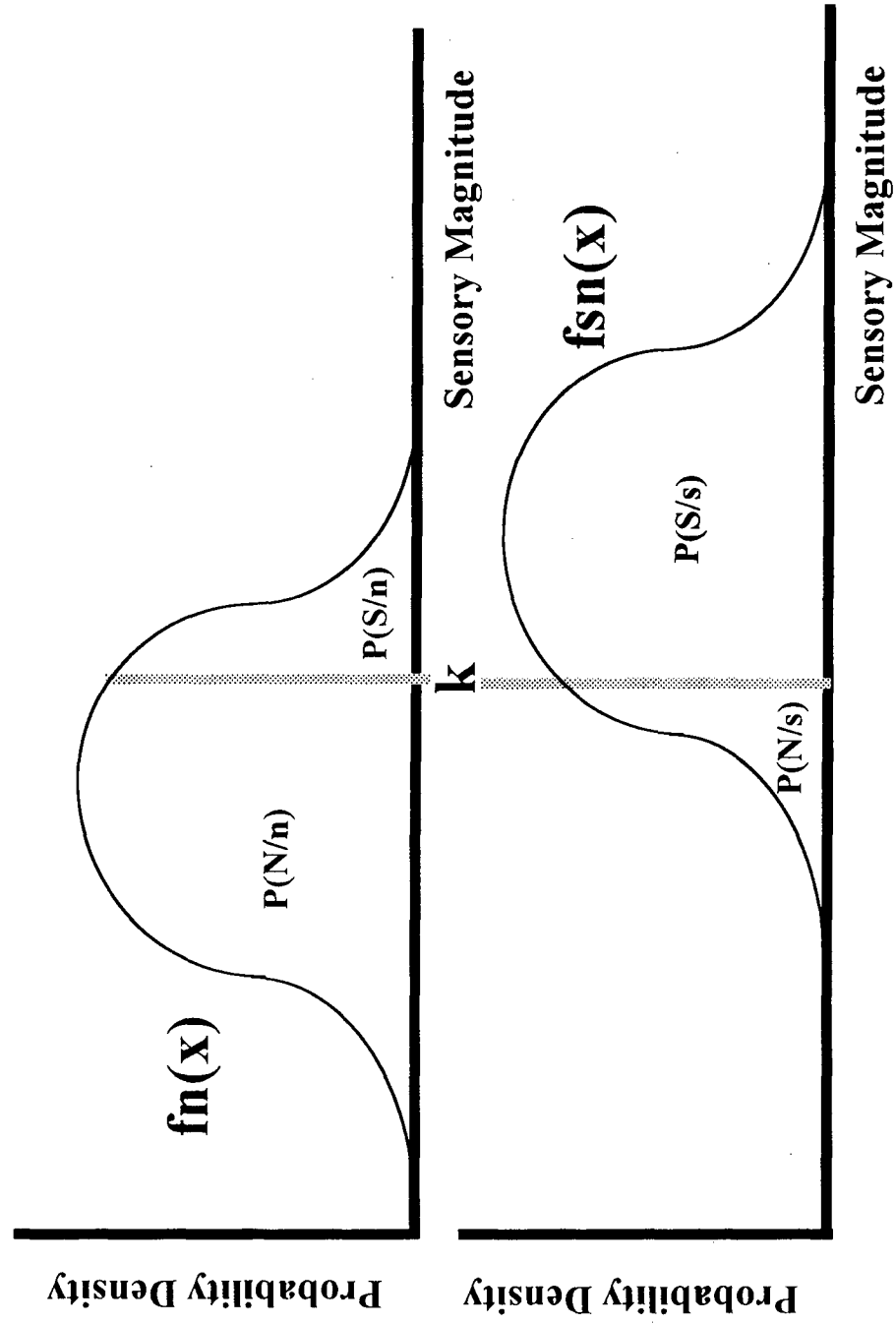


Figure 5. The four event regions of the decision space distributions.

If the stimulus is noise only (n) and the observer provides a correct “noise-only” response (N), he has *correctly rejected* the stimulus. The conditional probability estimate, $P(N/n)$, of this event is represented by the area under $f_n(x)$ to the left of value k . If the observer responds “signal plus noise” (S) to the noise-only stimulus, he has produced a *false alarm* decision. The false alarm estimate of conditional probability, $P(S/n)$, is the area under $f_n(x)$ to the right of line k and is referred to as the false alarm rate. If the stimulus is signal plus noise (s) and the observer responds correctly with “signal plus noise”, a *hit* is scored. The conditional probability estimate of a hit, $P(S/s)$, is the area under $f_{sn}(x)$ to the right of line k and is referred to as the hit rate. In the event of signal-plus-noise stimulus resulting in a “noise-only” response (N), the observer has *missed* the signal, and the estimate of conditional probability of misses, $P(N/s)$, is described by the area under $f_{sn}(x)$ to the left of line k . The operational target detection analogy should be obvious to the reader.

In order to quantify the performance capability of the observer in detecting and deciding upon the presence of a signal, some means of empirically determining d' must be employed for each set of stimuli. Receiver operating characteristic curves and estimates of d' can be derived through proper experimental procedures.

The Receiver Operating Characteristic Curve

The receiver operating characteristic curve defines the relationship between the false alarm rate and the hit rate for a given level of signal detectability. The ROC curve relates these two probabilities across the spectrum of operator response bias, and it facilitates analytical determination of d' . A formal presentation of the relationship between the ROC and the decision space distributions is beyond the scope of this research application. The interested reader will find a detailed description in Swets, 1964. Instead, an intuitive approach to ROC curves will be presented.

Figure 6 illustrates a typical ROC and the relationship to the decision space (Gescheider, 1976). The horizontal axis represents the false alarm rate (FAR)

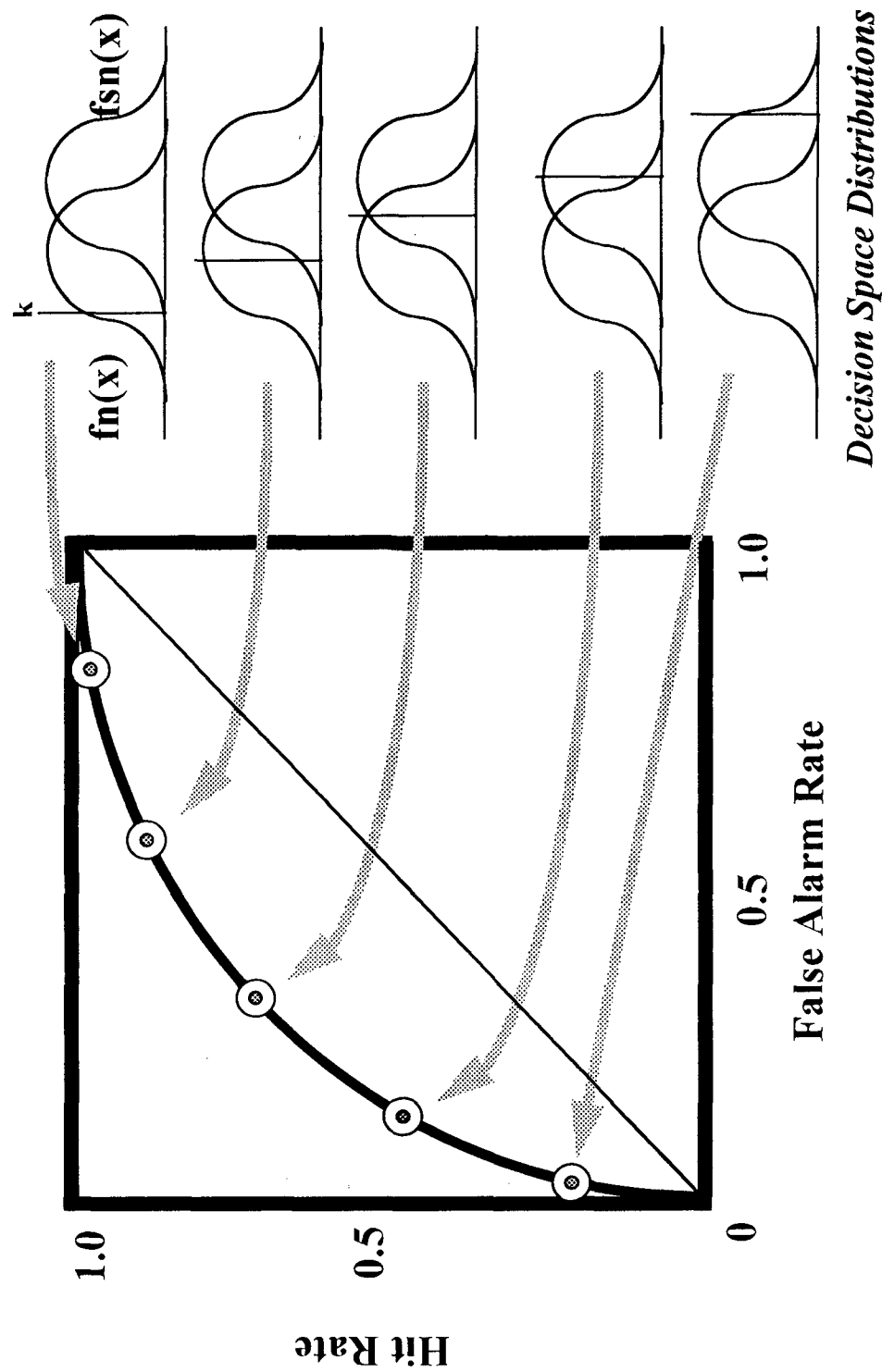


Figure 6. Relation between decision space distributions and ROC curve.

from the associated area under the $f_n(x)$ distribution. The vertical axis represents the hit rate from the associated area under the $f_{sn}(x)$ distribution. Points selected near the lower-left end of the curve are representative of very conservative decision criteria in which the criterion line is located far to the right in the decision space. This is consistent in that a low false alarm rate is coupled with a larger, but still small, hit rate. Points near the upper-right end of the curve indicate a high hit rate and high false alarm rate, consistent with very liberal decision criteria. Intermediate points appropriately represent the probabilities associated with intermediate bias values.

At any given point along the ROC curve, the decision bias value (B) is equal to the slope of the tangent drawn at that point (Swets, 1964). Figure 7 helps to intuitively grasp this point. Near the lower-left end of the curve, the slope of the tangent is great. Recalling that this region of the ROC is representative of a criterion line on the far right side of the decision space, we can see that the ratio $f_{sn}(x)/f_n(x)$ is also great. Conversely, near the upper-right end of the curve, the tangent slope is small. This is consistent with the small value of the bias ratio when the criterion line is near the left side of the decision space. Near the middle of the ROC curve, the tangent slope equals one, and this value agrees with the bias calculated at the intersection point of the two probability distributions where $f_{sn}(x)$ and $f_n(x)$ are equivalent.

Another important characteristic of the ROC is that the difference between the probability distribution means, d' , impacts the shape of the curve. While all ROC curves derived from normal distributions of equal variance have the same general form, the character of the curve within this general form is determined by d' . Figure 8 depicts a family of ROC curves of varying d' . Intuitively, a d' value of zero indicates perfectly overlapping probability distributions resulting in a constant bias-ratio value of one. This creates an ROC curve with a constant tangent slope of one. This ROC is the major diagonal in the unit square and can be considered as "chance" performance since adding signal to the noise provides no information and the detection system can only guess as to the presence of signal in

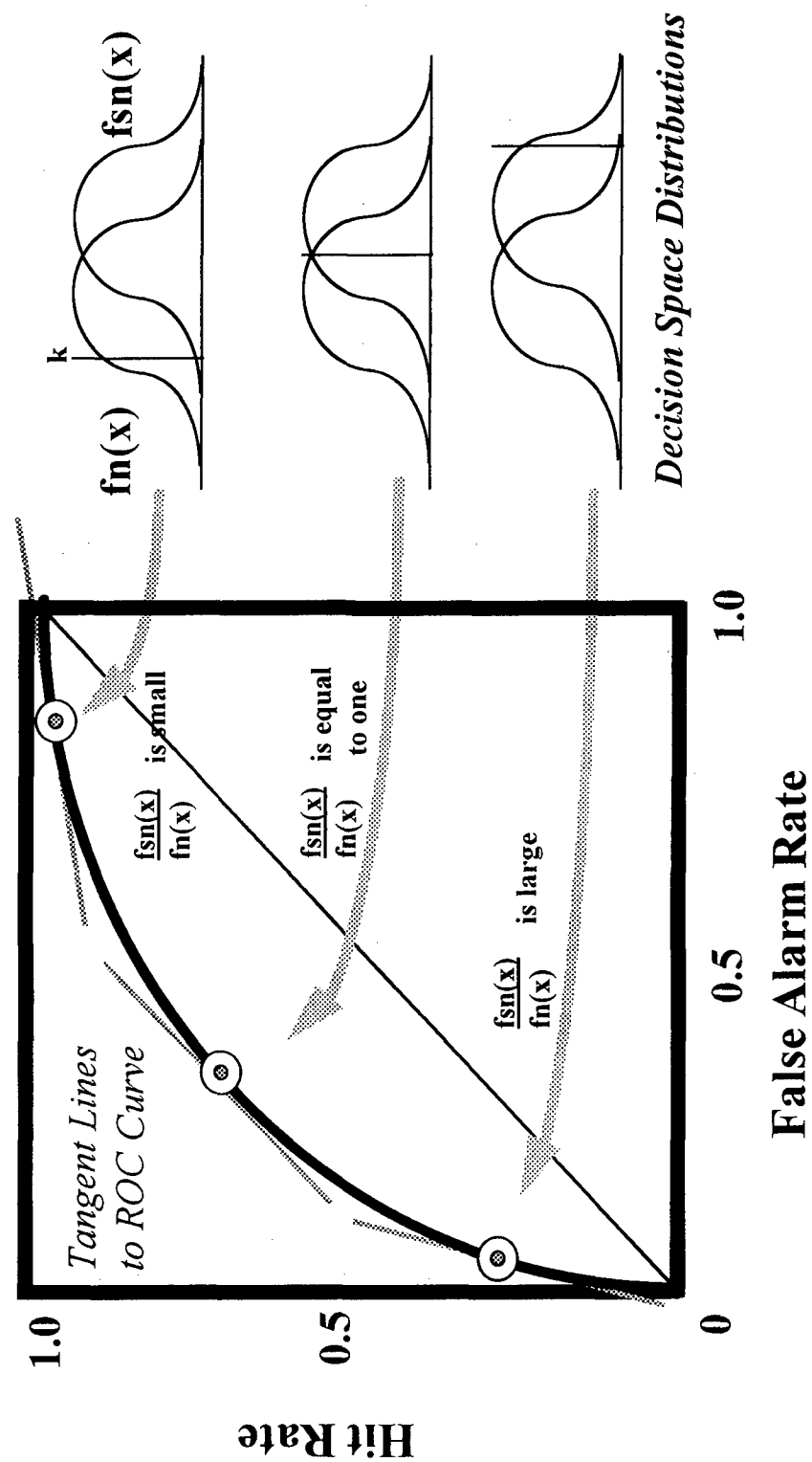


Figure 7. Relation between $f_{sn}(x)/f_n(x)$ ratio and the ROC curve.

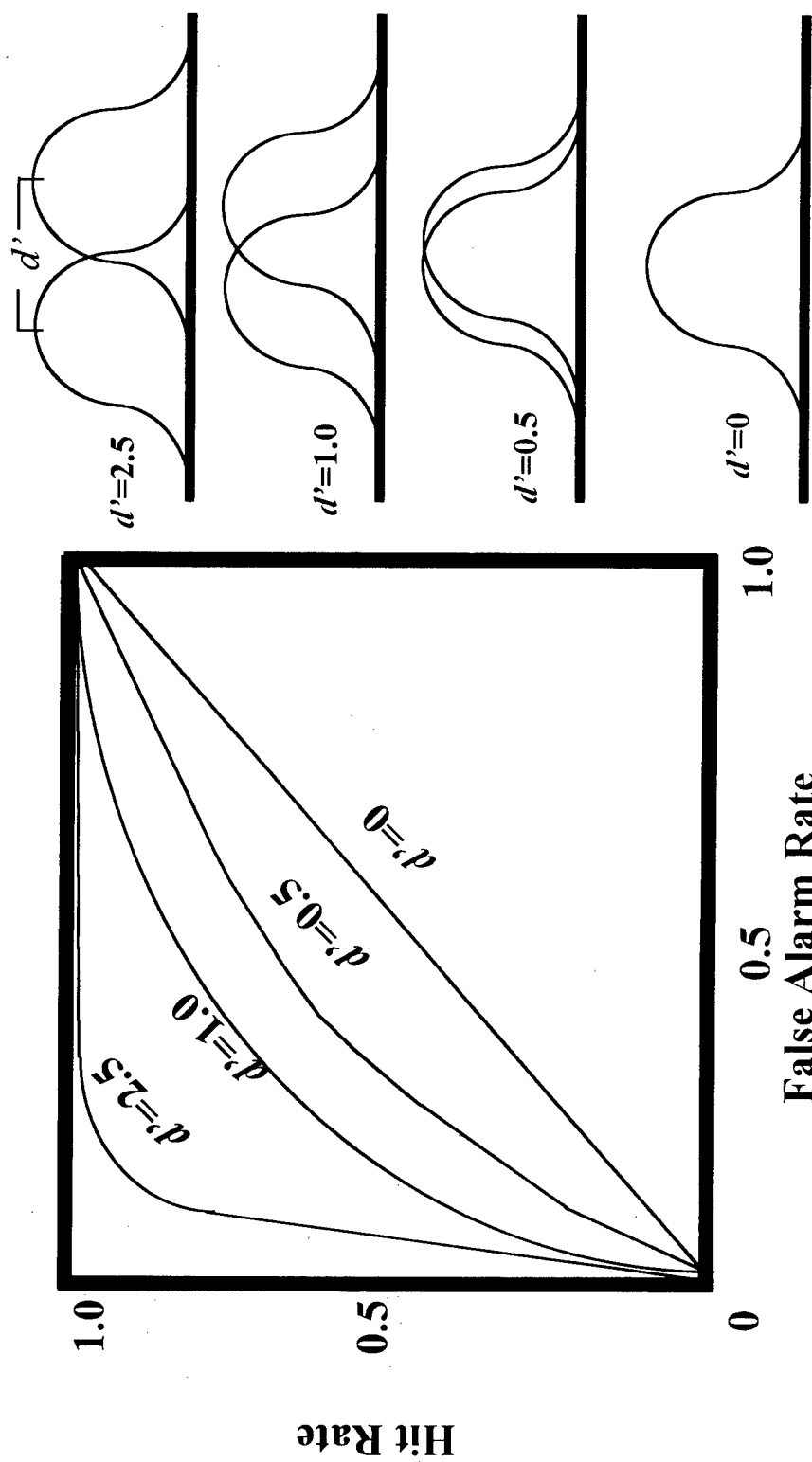


Figure 8. Family of ROC curves of varying d' .

the noise (Swets, 1964). In general, greater d' values move the central part of the curve toward the upper-left.

It should be noted that the exact form of ROC curves may vary depending upon the nature of the decision space probability distributions. The curves represented thus far are typical of the forms resulting from normal distributions of equal variance. Cases in which these distributions are not normal or do not have equal variance may dictate the use of alternatives to the d' metric.

Empirical ROC Curves and d' Metrics

Empirical data can be collected to generate ROC curves and to determine d' . An experimental procedure must be employed which requires subjects to make decisions about the presence or absence of signals in stimuli. Further, the experiment must induce changes in the decision criterion to generate different points in ROC space. Two well proven methods are available to the experimenter.

The first method is referred to as the "fixed-interval observation experiment," or the *Yes-No Experiment* (Egan, Schulman, and Greenberg, 1964). The concept is simple. Subjects are presented with numerous stimuli, one at a time. The subject must decide whether or not a signal is present and report his decision to the experimenter. The experimenter defines the detection task rules, or "payoff", prior to each experimental block. The subject must adjust his decision criterion based on the payoff and the goal of optimizing performance. The experimenter collects data points at several levels of bias and plots the hit rates and false alarm rates to generate the ROC curve and to calculate d' (Egan, Schulman, and Greenburg, 1964).

In the previously described operational targeting task, the experimenter could induce variable response bias by altering the rules of engagement and the payoff values of hits and false alarms. The afore mentioned infinite-munitions scenario, coupled with an operator score dependent upon only the number of target hits, produces data points of liberal bias to be plotted near the top-right in the ROC space. The finite munitions scenario, coupled with score reductions

proportional to the number of false alarms generated, would produce more conservative decisions. These data points are plotted near the lower-left in the ROC space. Intermediate rules and payoffs can be devised to generate middle points.

The yes-no experiment requires extremely large numbers of trials to achieve even moderate reliability. Egan (1975) has shown that during a series of observations the human is capable of adopting multiple criteria. This fact allows the conduct of *rating experiments* in lieu of the yes-no experiment, and a four-category rating scale produces reliability comparable to the yes-no experiment with about one-third the number of trials (Egan et al., 1964). In psychophysical experiments, the ratings are often referred to as “confidence ratings” intended to represent how certain a subject feels in judging the presence of the signal in a stimulus.

The rating experiment is similar to the yes-no experiment. Subjects are presented with a series of individual signal-plus-noise and noise-only stimuli after which each subject's response is recorded. Unlike the yes-no experiment, the response options are not binary. The subject must respond with a confidence rating corresponding to a predetermined set of criteria definitions. For example, a rating of 4 may correspond to a “yes” response under very conservative criteria, and may be described as a subject confidence report of “signal definitely present.” A rating of 3 would correspond to a “yes” response under intermediate criteria and relate to a confidence report of “signal probably present.” A rating of 2 can be thought of as a “no” response with lax criteria - “signal probably not present,” - or as a “yes” response with very liberal criteria - “maybe signal present.” The rating of 1 is then a conservative “no” - “signal definitely not present.” Several different rating schemes have been utilized in related research. The interested reader will find other examples in Astley, Taylor, Boggis, Asbury, and Wilson (1993), Gescheider (1976), Ozkaptan (1979), Swets (1964), and Wilson (1992).

Four response categories, such as those previously described, are defined by three criteria in the decision space. Figure 9 illustrates the four-response

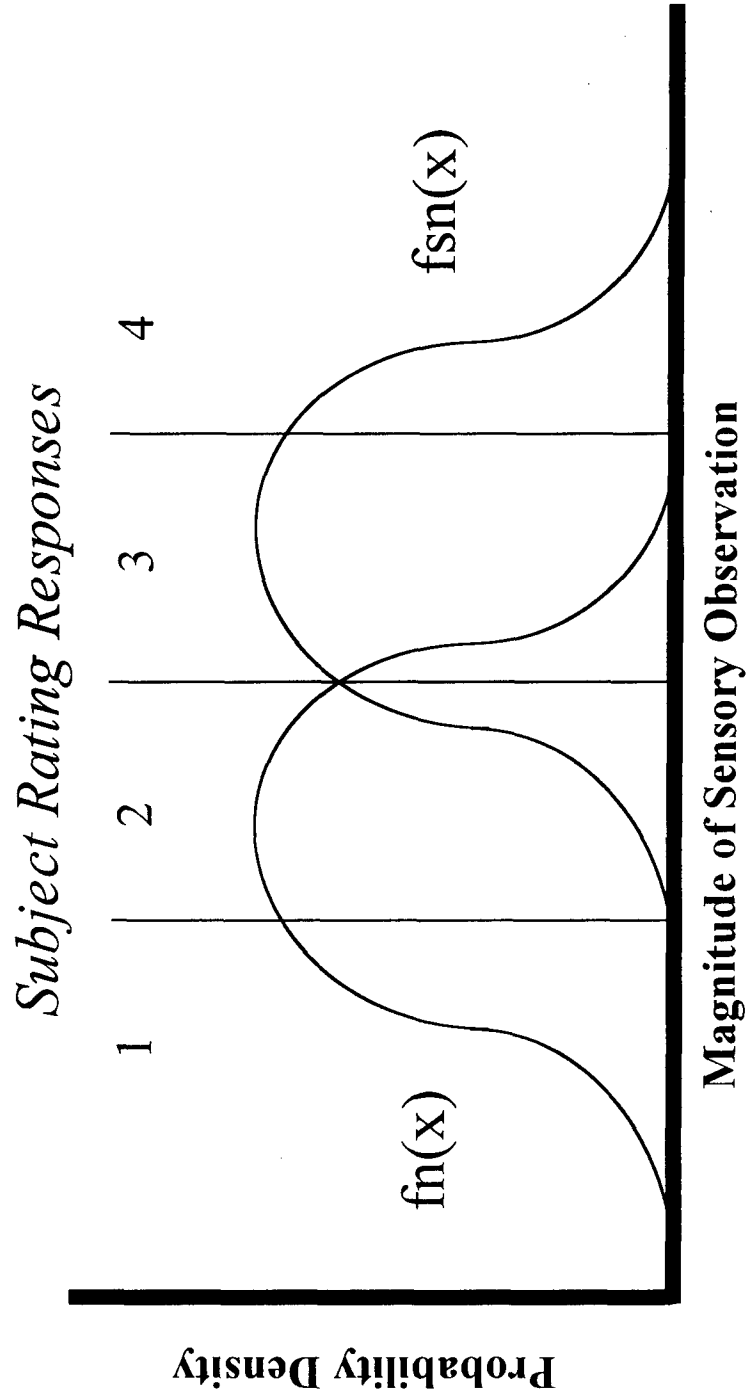


Figure 9. Relation of subject confidence reports to decision space criterion line.

scenario. Each response results from a unique set of sensory magnitudes. The sensory magnitude sets associated with each response are defined by the three criterion lines in the decision space. Examination of empirical subject responses reveals the positions of the criterion lines and allows the determination of three points on the ROC curve.

Analysis of the rating experiment is conducted by considering each criterion value separately as a single criterion dividing yes-no responses. First, the subject responses are analyzed with the assumption that the observer would say "yes" on those trials where a rating of 4 (signal definitely present) was provided, and all other responses are treated as if the subject said "no". Secondly, the subject responses are analyzed with the assumption that the observer said "yes" on trials where the rating of 3 (signal probably present) or higher was provided, and all other responses are equated to "no." The technique is applied to each response level for all noise-only and signal-plus-noise trials, allowing estimates of $P(S/s)$ and $P(S/n)$ for each criterion, and thus generating a point on the ROC curve for each criterion (Green and Swets, 1966). Table 1 illustrates the general stimulus-response matrix of the rating procedure using four rating responses.

Table 1. General stimulus-response matrix of estimates of conditional probability for four rating responses.

<u>Stimulus</u>	<u>Rating Response</u>			
	4	3	2	1
signal	$P(4/s)$	$P(3/s)$	$P(2/s)$	$P(1/s)$
noise	$P(4/n)$	$P(3/n)$	$P(2/n)$	$P(1/n)$

A specific example will illustrate the determination of hit rate and false alarm rate pairs from empirical data for ROC curve plots. Suppose that a total of 200 stimulus trials were presented to a subject, with 100 trials containing signal

plus noise and 100 trials containing noise only. Table 2 depicts response frequencies for a sample stimulus-response matrix.

Table 2. Sample stimulus-response frequency matrix for 200 trials.

<u>Stimulus</u>	<u>Rating Response</u>				<u>Total</u>
	4	3	2	1	
signal	50	25	15	10	100
noise	5	20	30	45	100

The conditional probability estimates for each cell of the matrix are calculated as proportions. Table 3 depicts the calculation of conditional probability estimates for each rating response and stimulus condition.

Table 3. Sample stimulus-response conditional probability estimates matrix for 200 trials.

<u>Stimulus</u>	<u>Rating Response</u>				<u>Total</u>
	4	3	2	1	
signal	.50	.25	.15	.10	100
noise	.05	.20	.30	.45	100

Cumulating the probability estimates of Table 3 generates conditional probability estimates representing the isolated examination of each criterion in the yes-no context. Table 4 depicts the cumulated probability estimates for the example.

Table 4. Cumulative conditional probability estimates matrix for four response ratings.

<u>Stimulus</u>	<u>Rating Response</u>			
	4	3	2	1
signal	.50	.75	.90	1.0
noise	.05	.25	.55	1.0

The conditional probability estimates of Table 4 are interpreted as hit rate and false alarm rate pairs to be plotted as points on the ROC curve in the unit square. The anchor points (0,0) and (1,1) are included in the plot representing theoretical criteria extremes.

The value d' is estimated using these same values from Table 4. The probability estimates are each transformed into z-scores. The difference between the z-scores of each pair of hit rate and false alarm rate values provides an estimate of d' . A more detailed description of this process follows in the *Method* section.

In classical applications of TSD, the number of signal-plus-noise stimuli, or hit opportunities, and the number of noise-only stimuli, or false alarm opportunities, is a known quantity. That is, the experimenter can control the number of stimuli which contain signal and which do not. This facilitates the calculation of the conditional probability estimates. However, it is possible to imagine stimulus scenarios in which the false alarm opportunities are not easily determined quantities. This is the case with the previously described operational scenario. Obviously, without a means of determining the denominator in the calculation of conditional probability, the probability is impossible to determine.

Hit - False Alarm Count Curves

Because ATR evaluation requires a dynamic video scene such as that described in the FRACTIL demonstration, and because evaluations with only a single target within a particular scene are not operationally relevant, dynamic video

scenes containing multiple target opportunities are employed. The ROC and d' would be good metrics to gauge the target detection performance of an ATR, but the unique stimulus (input scene) requirements prevent the use of traditional TSD techniques.

While the total number of target opportunities (signal-plus-noise stimuli) can be determined from a count of targets in a scene, the total number of false alarm opportunities (noise-only stimuli) is certainly questionable. How does the experimenter determine probabilities of false alarms if there is no easily determined denominator in the false alarm ratio? How can a total number of false alarm opportunities be determined in a dynamic video scene? It can be argued that there are nearly infinite opportunities for false alarms under these stimuli. At best, each pixel combination with a visual angle equivalent to that of a target's visual angle could be considered a false alarm opportunity. The possible combinations are extremely large, and the calculation of false alarm probability is meaningless.

These difficulties have led ATR evaluators to utilize absolute counts of false alarms rather than false alarm rates. The adopted community standard of evaluation plots hit rate against false alarm count (ATRWG 86-001, 1986). A typical hit-FAC curve is illustrated in Figure 10.

Most ATR devices provide "reports" at specific intervals, or frequency. A report consists of a listing of "objects of interest" (OI) within a scene which fit the ATR criteria for a target. Because different ATR devices operate at different frequencies, the false alarm count must be translated into a common format. A common treatment of this problem is to present the false alarm data as the average number of false alarms per ATR report for a given range bin. An ATR which reports only once within a range bin can then be compared with an ATR which reports on every video frame, or a frequency of 30 Hertz.

The hit-FAC curve is produced empirically from ATR reports. In the Wright Laboratory evaluation of three ATR devices, each device reported on multiple OI's and provided a numerical "confidence rating" for each. The rating is

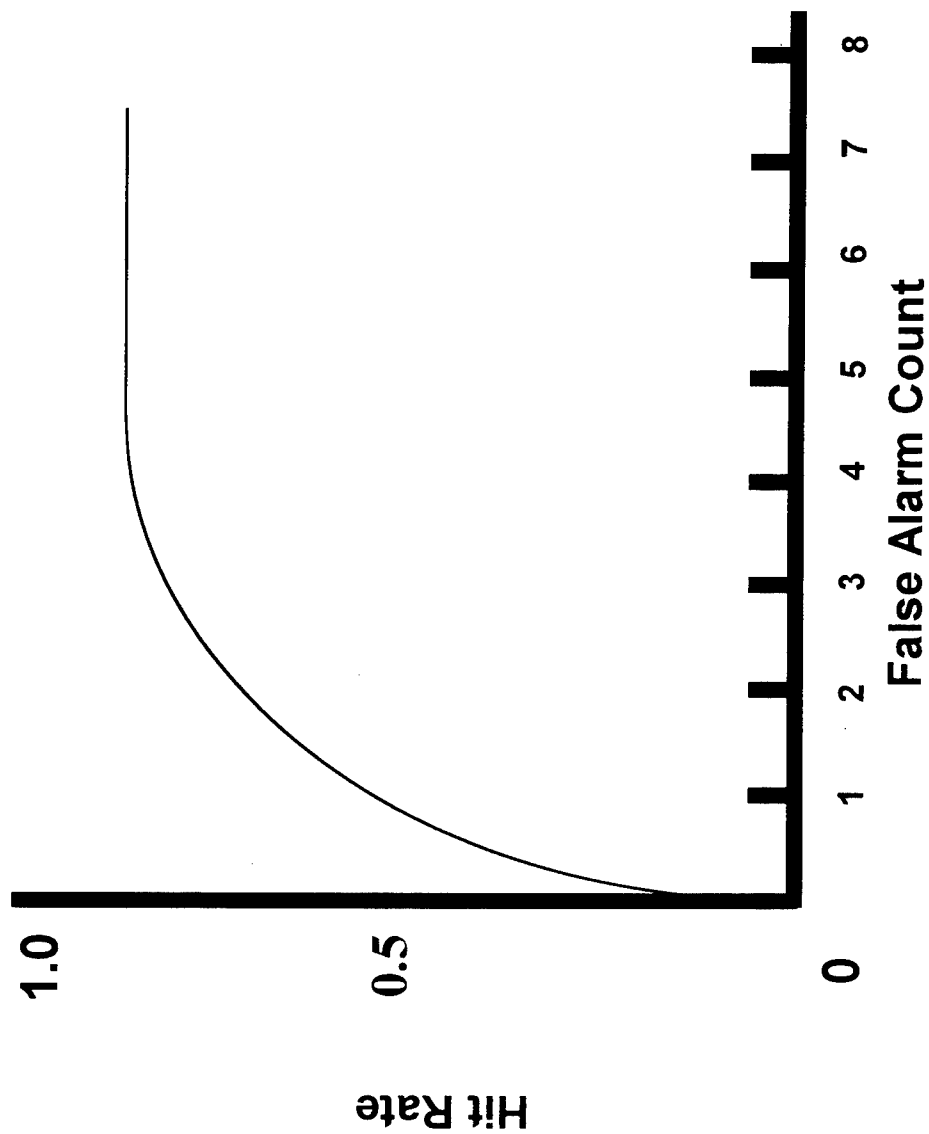


Figure 10. Typical Hit-FAC curve for ATR performance.

actually a mathematically determined value which represents the level of “agreement” between the thermal signature of the OI and electronically stored exemplar signatures (Clark, 1994, personal communication). The OI ratings were ranked in numerical order from highest confidence to lowest.

The hit-FAC curves were then generated through a graphical technique. Figure 11 provides an example of the generation of one range bin’s hit-FAC curve for a scenario in which five reports were made by an ATR within the bin. In this example, five targets were present in the video scene throughout the range bin.

The ordinate is scaled into n equal increments where n is the total number of hit opportunities across all reports within the range bin. Since there are five reports each with five possible hits (five targets), the vertical axis is partitioned into (5×5) twenty-five segments. An incremental cumulative plot of target hits will indicate the hit rate at any given false alarm count per report.

The abscissa is defined as false alarms per report. To estimate this value, each unit of false alarm on the axis is sub-partitioned by the total number of reports within the range bin. In the example, each unit of false alarm is partitioned into five segments since there are five total ATR reports in the range bin. Under this scheme, an incremental cumulative plot of false alarms will indicate mean false alarms per report.

The confidence ratings associated with OI’s are listed under each report header in the upper-right of Figure 11. Below the separate report listings, the OI ratings are ranked in descending order of confidence values across all five reports, and the hit or false alarm character of each report is noted (Only a partial ordering of the whole set is depicted for convenience). The hit-FAC curve is then plotted based upon this hit-false alarm character as summarized under the heading “Plot Actions.” Starting at the origin point, hits move the plot up by one increment. Similarly, false alarms move the plot to the right by one increment. As the OI list is reviewed from highest confidence value to lowest, the hit-FAC curve is generated (Clark, 1994, personal communication).

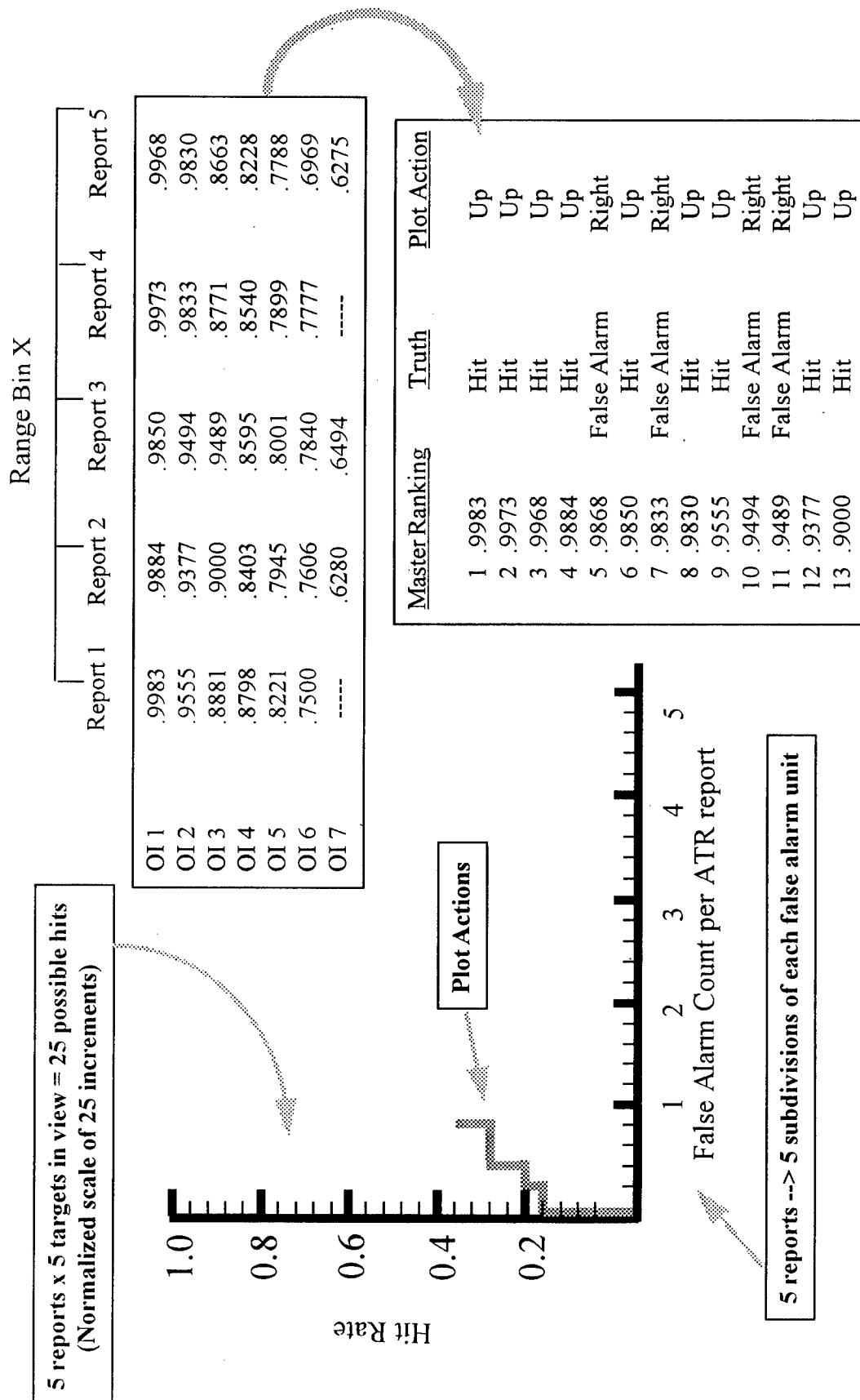


Figure 11. Example of Hit-FAC curve creation from empirical data.

The similarity to ROC curves is obvious. The highest ATR confidence values are plotted first (nearest the origin) just as the most conservative bias points are plotted for the ROC. Generally, these higher ratings consist of more hits than false alarms and produce a curve segment with a large tangent slope. Since lower ratings representative of more liberal bias are plotted later, they necessarily appear toward the upper-right of the curve space. Because they tend to produce more false alarms, they generally result in shallower tangent slopes. While it would be improper to equate these slopes to any conditions in the decision space distributions, the general shape and character of the resultant curve are similar to those of ROC's. Comparisons can be made between hit-FAC curves to gauge performance, but here the similarity to ROC's ends. No conclusions regarding the decision space can be made. No equivalent to the d' measure exists.

Comparisons can be made between human performance and ATR performance if hit-FAC curves are generated for human operators. Human hit-FAC curves can easily be created using the same video imagery employed in the ATR evaluations. Obviously, the human cannot react and report as quickly and frequently as the ATR's. However, a single report per range bin with multiple OI's can be recorded with proper experimental procedures.

The creation of ROC curves for the human in this task is also desired. The ROC's will establish the decision space distributions as a well defined reference for the human performance represented in the hit-FAC curves. Over time, a database which relates hit-FAC curves to empirically derived ROC's may serve as a bridge between the two metrics. The ROC's can also be used to establish the impact of target range and atmosphere-clutter. Further, the ROC curves will serve as a baseline of human performance for related follow-on studies.

RELATED RESEARCH

LANTIRN FRACTIL Research

The Air Force advanced technology demonstration known as FRACTIL was conducted in 1993, and it was the activity from which the previously described operational scenario and video stimuli were derived. During the demonstration, Turner, Purvis, O'Hair, Malek, and Reynolds (in press) conducted an operationally oriented performance assessment of a prototype ATR system embedded in the current production LANTIRN targeting system. The assessment compared the pilot-system targeting performance of the current LANTIRN configuration with the configuration including the ATR.

The assessment encompassed twenty-nine flights, each with four target passes dedicated to the comparison. Two of the four passes were flown without employing the ATR and two were flown fully utilizing the ATR. The pilot was provided with target array coordinates of variable accuracy to induce an area of uncertainty for target search.

For the non-ATR passes, the pilot was instructed to use his FLIR sensor to locate the target array and then sequentially designate and simulate a missile launch against individual target array elements. He was instructed to perform as many simulated launches as possible during the fly-over attack as rapidly as he could.

During the ATR-assisted passes, the pilot's duties were to initiate the ATR in search mode and, once the ATR had detected the array, follow its recommendations explicitly. The ATR automatically conducted sequential target designations in order of confidence rating values. After each designation the ATR waited for the pilot to respond with a "launch" command or a "disregard" command. Following the ATR recommendations, the pilot responded with launch commands and fired as many missiles as possible as rapidly as possible.

Target designation actions and simulated missile launch actions were recorded on the targeting system FLIR video recorder. These actions and other flight-related parameters were portrayed through computer generated symbols and alphanumerics displayed over the FLIR video scene. Slant-range to target was displayed and facilitated the grouping of targeting actions within range bins. Additional information was recorded digitally and correlated through a common time reference.

The evaluation compared numbers of correct target designations, numbers of incorrect target designations, numbers of simulated missile launches achieved, the stand-off range of targeting events, and other metrics associated with performance and pilot workload. Subjective pilot data were also collected to supplement the objective results. Subjective data included the Subjective Workload Assessment Technique (SWAT) reports for estimating pilot workload under the different conditions, and post-flight structured interviews were conducted to collect pilot comments. The specific results of the FRACTIL assessment are classified by the Department of Defense and are not reported in this document.

The FRACTIL demonstration generated only trend information. In field research sufficient numbers of trials and adequate experimental control are often difficult to achieve. This was certainly the case during FRACTIL, where restraints on resources and competition among multiple data collection requirements prevented the conduct of a more sound comparison. These limitations were partly the impetus for subsequent laboratory evaluations of ATR and human performance in the target detection task.

Wright Laboratory ATR Evaluations

Following the FRACTIL demonstration, Clark and Westercamp (1994, personal communication) initiated a laboratory effort to quantify the performance of three different ATR devices designed by three different defense contractors. The digital video of various target arrays collected during FRACTIL was used as

the input data for these evaluations. Hit-FAC curves have been produced for eight range bins for each device using the previously described method.

The three ATRs provided reports on multiple OIs, each at a different reporting frequency. Device A reported on every video frame for a frequency of 30 Hertz. Device B reported at 3 Hertz, and device C reported only once per second. These frequencies roughly equate to 150 reports per range bin, 15 reports per range bin, and 5 reports per range bin, respectively.

Each of the three ATRs employed a unique paradigm as described by Zelnio (1987). Device A operated with the *Prescreen, Segment, Classify* (PSC) paradigm. Device B employed a *Matched Filter* paradigm, and device C used the most sophisticated paradigm to date, the *Model-based Vision* (MBV) paradigm.

Although each ATR functioned differently, hit-FAC curves were generated to describe the performance of each. Ultimately, it is these ATR results which must be employed in direct comparisons with human performance results to gauge the effectiveness of each ATR. The current research seeks to derive human performance measures to facilitate the comparison.

Barnes

The Naval Weapons Center conducted a target acquisition study in 1977 which primarily examined the impact of display size on target detection performance by human operators. Additional factors were examined by Barnes (1978) which provide insight into the image and target characteristics that provide salient cues to the observer searching for targets.

Dynamic video imagery was used for the stimulus. The video was generated from a realistic terrain model on which model targets (tanks and trucks) were placed in variable configurations. The imagery was collected such that it was similar to that produced by a forward looking optical sensor flying over the terrain. A fixed altitude of 3200 feet above ground level (AGL) was simulated with a 2-degree field of view. The background clutter was described as "medium European", indicating substantial trees, brush, and other variable features. The

simulated airspeed was 363 knots. The video segments were displayed on a 300-line resolution TV monitor.

Independent variables included target visual angle, number of targets present in the scene, display contrast, and target configuration. Three levels of target visual angle were employed (7, 27, and 47 minutes of arc), although no explanation is provided of how these values were determined with the dynamic video display. The assumption is that these are mean values of visual angle for each presentation segment. Three levels of target numbers were presented (one, four, and seven) with two contrast levels (light - 27% and dark - 48%). Two levels of the target configuration variable were examined (linear and random configurations). Display size was also varied, but target visual angle was held at each of the three prescribed levels.

The basic procedure presented the subjects with segments of video representing 4 nautical mile (nm) stretches of terrain. After each trial, a tone sounded to alert the subject of the initiation of the next trial and to induce a forced-choice response. The subject depressed a "YES" button if he felt that a target was present in the scene, or the "NO" button if he believed that no target was present.

The within-subject design replicated each cell of the $3 \times 3 \times 3 \times 2 \times 2$ full factorial experiment four times, resulting in 432 data points per subject. A 300-trial practice session preceded data collection, and rest periods were allowed during data collection.

The results indicated that correct detection improved substantially when more than one target appeared in the scene. However, correct detection was nearly identical for the four and seven target configurations, suggesting a performance plateau with four or more targets in the scene. The linear configuration of targets aided detection only when the contrast variable provided target luminances similar to the luminance of prominent background clutter objects. The conclusion is that the linear array highlighted the targets. A performance plateau was also experienced for the visual angles of 27 and 47

minutes of arc. Barnes concluded that visual angles greater than approximately 27 degrees would not impact performance significantly. He also concluded that the monitor display size has no impact on detection performance for equivalent target visual angles.

In a prescreening experiment, Barnes determined that the display factors described above were the most significant sources of detection variance. Another factor in the prescreening experiment was display resolution. Imagery was displayed on the 300-line TV as well as on a 175-line TV. While no specific quantification of this variable's impact is provided by Barnes, it is described as having a minor impact on performance.

Barnes' research is significant to the current research in that the target detection tasks are similar. Some of the procedural techniques employed in the current research are derived from Barnes' procedures for presenting dynamic video stimuli. Most significantly, Barnes' results regarding imagery resolution are relevant since the imagery to be employed in the current research is of slightly lower resolution than that available in operational aircraft.

Ozkaptan

In 1979, Ozkaptan performed an evaluation of the utility of TSD for target acquisition studies for the US Army. He simulated helicopter pop-up maneuvers using static images of terrain scenes with 30-sec subject observation times. Terrain scenes contained either a single military tank target or no target.

A yes-no TSD technique was employed with three levels of instruction to alter response bias. Instructions were tailored to emphasize accuracy (conservative bias), neutrality (neutral bias), and speed (liberal bias). Two levels of image contrast were presented with four different background scenes in a modified Latin Square design. Each subject was exposed to 30 trials for each experimental condition.

Ozkaptan concluded that the instructional set is important in determining aviator performance during target detection tasks. The allotted response time (as

induced by variable instructions) had a significant effect on detection performance as measured with the d' metric. He also noted that the signal detection parameters could be employed to "remove the effects of different instructional sets." He postulates that the TSD parameters could be averaged over subjects and used as dependent variables to compare the effectiveness of different sensor systems, without the confounding effects of response bias. He further acknowledges that these parameters could also be used to remove the effects of non-system-related factors of sensitivity and bias from operationally relevant measures.

Ozkaptan set a precedent by using TSD in the evaluation of military target detection performance of aviators. His suggestion of averaging TSD measures over subjects for comparisons of different sensor systems applies as well in the comparison of different sensing environments using a single sensor. This is one technique which is employed in the current research to evaluate performance in variable atmospheric and clutter conditions.

Astley, Taylor, Boggis, Asbury, and Wilson

In 1993, Astley et al., encountered a visual stimulus scenario similar to the ATR scenario. These researchers were attempting to quantify performance of machine-assisted analysis of medical imagery, specifically mammograms. Two different ATR-like devices, or "cue generators," were employed to assist in the detection of microcalcifications which are early indicators of breast cancer. The imagery used for stimuli typically contained multiple microcalcifications in clusters and in isolated occurrences. The problem here, as with the FRACTIL imagery, is that multiple targets must be presented simultaneously and there is no ready means of quantifying the opportunity for false alarms.

The researchers devised a novel solution. After each mammogram had been "truthed" by expert radiologists (the microcalcifications were identified and annotated), regions of interest (ROI's) were created which fractionized the image into various polygonal areas. The ROI's were defined so that each contained at least one microcalcification and so that no microcalcification appeared outside of

the set of ROI's. Within each ROI, a circular area was defined about each microcalcification marker and referred to as a "disk." The disk radius was defined as 16 pixels, or 0.8 millimeter. The disk area represented an area of correct microcalcification detection, that is, a "hit" area.

The two cue generators were then independently applied to each image. Like an ATR, the cue generator produced confidence reports associated with each object of interest that it identified within the overall image. The single confidence report with the greatest value within each ROI was considered to be the "on target" response from the cue generators for that ROI. All other reports within an ROI were ignored.

Each ROI was considered an independent stimulus presentation. Maximum value reports were considered "hits" if the reported object was within the predetermined disk area, while maximum value reports for objects outside of the disk area were scored as "false alarms." Since only a single report was considered for each ROI, the opportunity for false alarms was fixed at a value equivalent to the number of ROI's presented. Similarly, the hit opportunity is also fixed at the same quantity.

The response threshold of the cue generators was controllable and was adjusted across the entire available range. This effectively altered response bias across the decision space of the cue generator and allowed the generation of ROC curves with the *yes-no* technique. ROC curves were plotted and used to compare the performance of the two devices. The same technique could be employed to compare the performance to that of student radiologists in detecting microcalcifications.

The technique employed by Astley et al., did not have the requirement of addressing dynamic video imagery. The ROI's and disks would be difficult to accurately define under dynamic conditions. Human subjects would have a difficult time providing numerous target detections and high-resolution confidence ratings in the brief time allowed by the operational scenario. However, a

modification of their technique may be applicable to the dynamic situation employing human subjects.

SUMMARY

Developing automated technology to assist military pilots with target detection tasks is a defense department priority. In order to gauge the value of newly developed ATR technologies, target detection performance must be compared with human performance of the same task. To achieve this, an accurate and detailed baseline of human performance must be established, and a common means of measuring performance must be developed. New and innovative evaluation techniques will be required to bridge the functional gap between human operators and electronic systems.

The Theory of Signal Detection and its associated metrics can be applied to establish a human performance baseline. The TSD metrics offer a well established means of comparison between experimental conditions, independent of operator bias. The TSD description of the decision space distributions provides an anchor for the application of other target detection performance measures, such as the ATR hit-FAC curve technique.

Because the hit-FAC curve technique is a standard approach to evaluating ATR's in the laboratory, it is a convenient measure to employ in comparing human performance to ATR performance. Three ATR devices have been tested using previously collected operational infrared video imagery, and the associated hit-FAC curves have been plotted. The creation of hit-FAC plots for human operators observing the same video stimuli will provide a good means of comparison between the human and ATR performance. Anchoring the human hit-FAC plots to a well defined decision space using TSD will facilitate condition comparisons in this study and in follow-on efforts.

The preservation of operational relevance is highly desirable in evaluating military systems performance. In comparisons of human and ATR performance, the use of dynamic operational stimuli and its inherent time limitations are the

minimum experimental constraints for preserving operational relevance in the laboratory. However, dynamic video stimuli with multiple imbedded targets presents some difficult problems for the employment of TSD techniques. Innovative experimental procedures are required to conduct a proper TSD experiment while preserving the required operational relevance.

Past research offers several techniques which provide direction in the creation of an experimental procedure to resolve the difficulties embodied in the human-ATR comparison task. Turner et al., (in press) defined the operational task, while Clark et al., (1994, personal communication) described the ATR evaluation process. Barnes (1978) provided valuable guidance on the use of dynamic stimuli as well as factors which may impact target detection performance. Ozkaptan (1979) summarized the value of TSD to target acquisition studies. Astley et al., (1993) defined a novel approach to dealing with the multiple-target problem.

Combining aspects of each of these past research efforts with newly conceived experimental techniques can result in an accurate and operationally relevant assessment of human target detection performance which lends itself to comparison with ATR performance of the same task.

OBJECTIVES

The general focus of this research concerned the ability of military aviators to detect ground-based targets imbedded in a dynamic video presentation of cluttered terrain. Specifically, the research sought to quantify the target detection performance of a human operator using a forward looking infrared sensor in low-level flight.

The research objectives were twofold. First, the intent was to derive measures which accurately described the detection task and which allowed comparison between experimental conditions and among related follow-on research results. The Theory of Signal Detection was employed to generate ROC curves for human subjects performing the operational target detection task. The d' metric was calculated for performance comparisons. Comparisons examined performance variations among target RANGE BINS and between ATMOSPHERE-CLUTTER (ATM-CLUT) conditions.

The second objective was to validate a technique for generating measures which would allow the direct comparison of human performance with ATR performance of the same detection task. Specifically, hit-FAC curves were generated for human subjects and a correlation between the TSD metrics and the hit-FAC metrics was sought. The TSD d'_e metric was correlated with the hit-FAC curve hit rates associated with predetermined values of false alarm count. Ultimately, bin-by-bin comparisons between human and ATR performance will be made using hit-FAC curves in a follow-on research effort.

Assumptions

In this experiment, subjects were presented with segments of FLIR video representative of a partial target ingress pass. During a specified RANGE BIN, subjects detected and pointed out the location of potential targets as expediently as

possible in coarse order of confidence. These data facilitated the creation of hit-FAC curves for each RANGE BIN. During the presentation of a subsequent RANGE BIN, subjects observed a sequence of circumscribed regions of interest. For each region the subject provided a confidence rating for the presence or absence of a target within the prescribed region of the video image. These data facilitated ROC curve generation and the calculation of d' . It was assumed that the decision space distributions of noise only and signal plus noise are normally distributed and of equal variance, promoting the use of d' .

The time associated with each RANGE BIN was brief, approximately 5 sec. The subjects observed a video segment recorded prior to and leading up to the specified RANGE BINS to acclimate themselves to each trial scenario. Quick action was required of the subjects in order to provide the necessary data. The assumption was that this swift reaction criterion was representative of the operational task. This was substantiated by the fact that the video presentation was "real time" and identical to the cockpit chronology.

Obviously, laboratory subjects did not have all of the distractions inherent in a real cockpit scenario. Artificial secondary tasks would have prevented the subjects from providing all of the information necessary to complete the research. Thus, the performance recorded is likely to be representative of the best performance possible in a similar real-world task. This assumption may be substantiated through an informal comparison of laboratory results with the results of the FRACTIL performance evaluation.

A third assumption was that the human can parallel process visual information to some extent. That is, the subjects could perceive multiple targets simultaneously and then react to them in a coarse order of confidence. The parallel visual processing assumption has been substantiated by numerous basic research studies (e.g. Beck, 1993; Overington, 1976).

Hypotheses

The first set of hypotheses address human target detection performance among the various experimental conditions. Based upon observations in the FRACTIL technology demonstration and previous target detection research, it is hypothesized that performance will degrade significantly with increased range to target under each of two atmosphere-clutter conditions. The two atmosphere-clutter conditions were Edwards AFB and Eglin AFB. However, the performance degradation with range under the Eglin conditions is expected to be more severe than that under Edwards conditions due to the deleterious effects of humidity on the IR image. At closer ranges, detection performance is expected to be nearly equivalent for the two atmosphere-clutter conditions since the impact of humidity on the IR image is less severe at shorter ranges, and ground clutter should have a lessened impact for targets of larger visual angle.

Additional hypotheses concern the attempt to correlate hit-FAC metrics with TSD metrics in order to validate the human hit-FAC technique. The hypothesis is that a linear relationship exists between the d' metric and the hit rate metric extracted from hit-FAC curves at preselected values of false alarm count. This necessitated that performance trends will be noted in the hit-FAC hit rate values similar to those trends exposed by the d' metric.

METHOD

Experimental Design

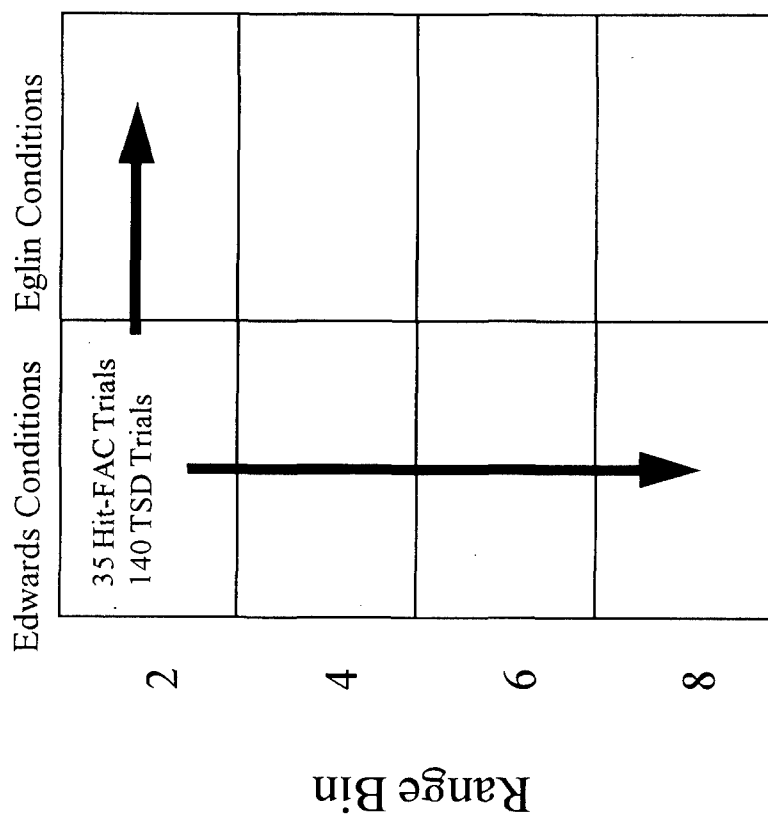
The objectives dictated collection of two separate data sets which are related but not directly comparable. The procedures for collecting the two data sets were necessarily different, but the data were obtained in a single, combined experimental design. For analysis purposes, the two data sets were considered as two separate experiments.

The independent variables were slant range to target in terms of one kilometer RANGE BINS, and combined atmosphere-clutter conditions. Four levels of RANGE BIN were examined and two levels of ATMOSPHERE-CLUTTER were examined. The RANGE BIN variable levels were identified as RANGE BIN 2, 4, 6, and 8, representative of the closest range to target, in kilometers, for each RANGE BIN studied. The ATMOSPHERE-CLUTTER levels were those of Edwards and Eglin. The dependent variables were the d' metric derived from TSD techniques and hit rate derived from the hit-FAC curve at a predetermined value of false alarm count.

The subjects observed 35 video segments per condition for each of the two data collection techniques, resulting in 280 RANGE BIN presentations per technique and 560 total presentations. For hit-FAC production, each RANGE BIN presentation was considered a single trial. Four individual trials were imbedded within each RANGE BIN presentation for TSD reports, resulting in 1120 total TSD trials.

For the hit-FAC data set and the TSD data set, the data were analyzed as a 4 X 2 factorial, within-subject, repeated-measures design (Figure 12). The hit-FAC curves were analyzed to determine each subject's hit rate for a false alarm count of two. In a preview of the subject data, a false alarm count of two

Atmosphere-Clutter



12 Subjects

TOTALS per subject:
280 Hit-FAC Trials
1120 TSD Trials

Figure 12. 4 x 2 factorial, repeated measures design.

seemed to be the value near which a majority of the hit-FAC curves began to plateau. Since this region of the curve depicts the most desirable operating performance (maximum hit rate for minimal false alarms), the value of two was selected as the false alarm count from which all subject hit rates would be derived. These hit rates were entered as a repeated measure and the TSD d' values were entered as a repeated measure.

Subjects

Twelve subjects participated in this experiment. Ten subjects were current or former military members with infrared imagery experience, or research and test engineers familiar with viewing operational infrared imagery. Two subjects were novices with no previous infrared imagery experience. Ten males and two females comprised the group. All subjects' normal or corrected vision met the following criteria: acuity of 20/20 Snellen equivalent or better as measured with a standard Snellen acuity chart; and normal contrast sensitivity tested with a Vector Vision CSV-1000 Contrast Sensitivity Tester.

Apparatus

The experimental apparatus represented a unique combination of video and computer technologies which facilitated the creation of stimuli and the cataloging of response data. The apparatus was designed to provide the subject with a very simple and ecological means of reporting -- touching the stimulus screen with a pointing stick (pencil). The touch responses were logged electronically and compared with truth data to determine hits and false alarms for generating both hit-FAC curves and ROC curves. A block diagram of the apparatus is presented in Figure 13.

The subjects viewed the stimuli on a Panasonic model WV950 monochromatic, high resolution monitor. This monitor rested upon a Touchmate touch screen device which used multiple pressure sensors to accurately determine the position of a screen touch. No intervening screen or other touch-sensitive

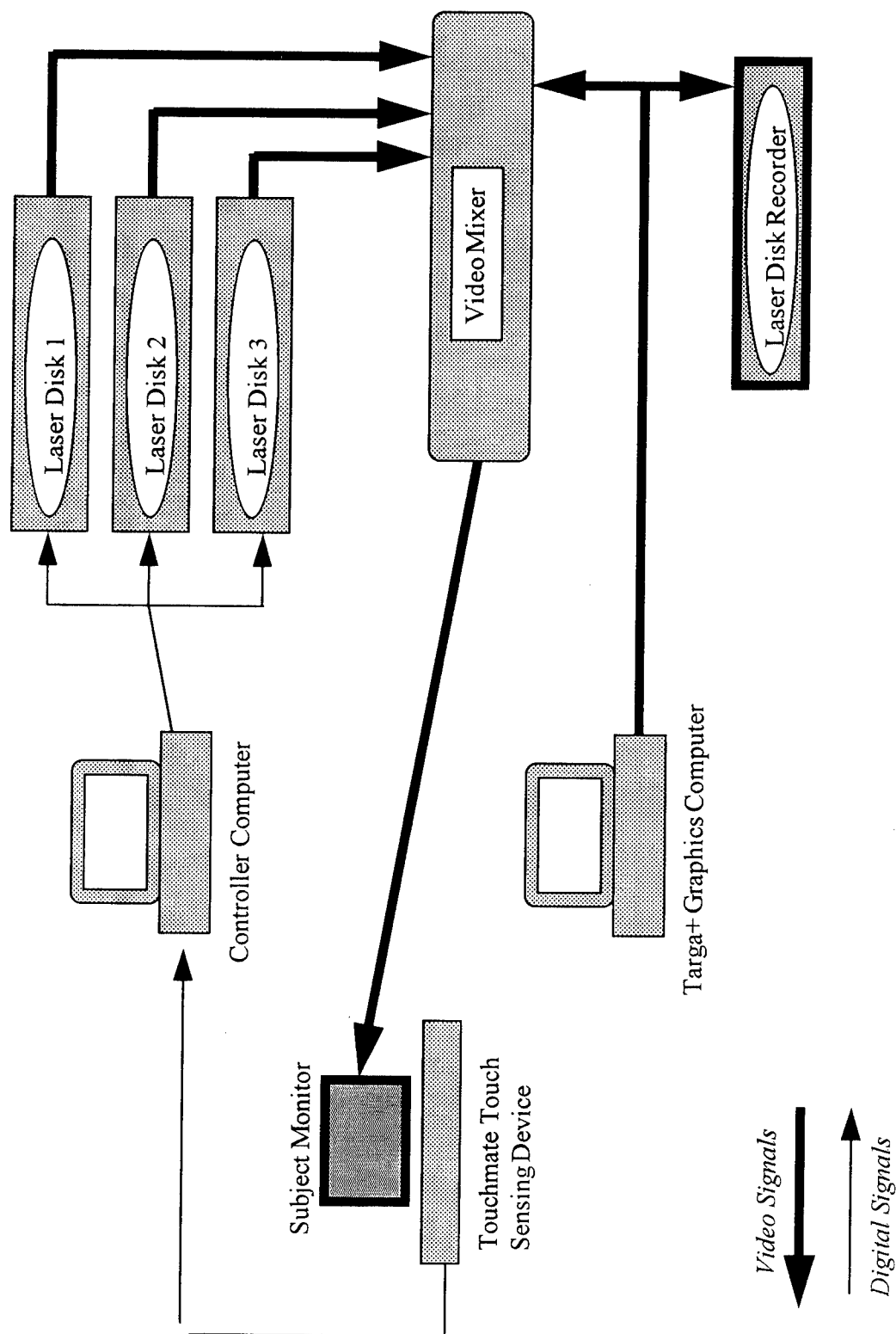


Figure 13. Block diagram of experimental apparatus .

plane was required between the subject and the presentation monitor, thus, no image degradation resulted from the employment of the touch screen device.

The Touchmate device acted as a touch-screen "mouse" and provided touch coordinates via RS232 connection to a 486 desktop computer. Screen touches were electronically logged and were compared with a database of predetermined screen coordinates representative of true target locations or a TSD rating value, depending upon the type of response provided.

The desktop computer controlled three Panasonic model TQ3032F video laser disk players. Each laser disk player could be individually selected by the computer to play precise segments of video imagery as identified by an input file of video frame numbers. The output signals of the three laser disk players were routed through a video mixing device which drove the monitor resting upon the Touchmate touch sensor. The result of this configuration was that the touch-sensitive monitor could display specified video segments from any one of the three laser disk players, and it could display combined video signals from two of the three laser disk players. The chronology and combinations of video presentations were dictated by the computer input file.

A 386 desktop computer was augmented with a special video effects board called TARGA+. The TARGA+ board generated user-defined graphics and output video signals which could be displayed on a standard video monitor. The effects board provided very accurate positioning of computer-generated graphics on the video monitor. With the previously described video mixing device, the TARGA+ computer-generated graphics could be displayed simultaneously with the laser disk player video signal. This allowed an overlay of computer graphics onto standard video presentations.

A Panasonic video laser disk recorder was used to transfer FRACTIL video tape segments onto laser disk platters. This recorder was also used to record graphics created with the TARGA+ computer system. Video signals from the laser disk containing the computer graphics were mixed with video signals from laser disks containing infrared imagery in a combined presentation.

Stimuli were presented on a monitor which reproduced the display size inherent in the F-16 LANTIRN system, a 5-inch square display. The display brightness and contrast controls were fixed at values which produced the highest contrast between the overall scene luminance and that of the individual targets as measured by a spot photometer. The video was displayed on a high resolution monitor which, when combined with the resolution of the video recording, produced a displayed resolution approximately 4/5 that of the actual LANTIRN system display. This resolution limitation was inherent in the recording medium, and the resolution achieved in this experiment was the highest possible with the recorded imagery outside the aircraft system.¹ It is important to note that the overall resolution was maintained at approximately 400 video lines vertical. In light of the results of Barnes (1978), this small reduction in resolution was not considered to be a significant factor.

Stimuli

The stimuli were selected infrared video segments of operational target ingress passes recorded during the FRACTIL technology demonstration. Each video segment presented a contiguous section of recorded video for the evaluation of two RANGE BINS. In total, four RANGE BINS were identified for evaluation: bin 8 (video imagery from 9 to 8 kilometers range to target), bin 6 (7 to 6 kilometers range to target), bin 4 (5 to 4 kilometers range to target) and bin 2 (3 to 2 kilometers range to target).

The video segments were selected from a superset of the FRACTIL video imagery. Segments were prescreened for suitability in this experiment. Segments were selected which exhibited little lateral motion of the IR sensor and little scene jitter or other visible anomalies. Consideration was given to maximizing the variability of target arrays and geographic location to minimize learning effects.

¹ The aircraft LANTIRN display is derived directly from the IR sensor and presents 480 lines vertical in the cockpit. The recording devices employed during the FRACTIL demonstration were not capable of preserving the image resolution at that level.

The most desirable video segments depicted a smooth, straight-in approach to the target arrays with little or no noticeable video defects or sensor movement. Some imagery was necessarily selected which contained undesirable characteristics due to limited available imagery. However, the impact of these flaws was minimized by employing these video segments such that few undesirable effects occurred during the actual presentation of the RANGE BIN for evaluation, but rather occurred before or after the RANGE BIN.

RANGE BINS were defined on the imagery through the use of an electronic signal recorded on the audio track of the original video tapes. An Integrated Range Instrumentation Group - B (IRIG-B) signal was recorded on the audio track defining an aircraft coordinated time code. This time code was cross referenced with aircraft range to target as recorded by the aircraft navigational systems in a separate time-range data base. The experimenter reviewed each video segment while observing the IRIG-B time code and annotated the video frame numbers associated with the beginning and end of each RANGE BIN of interest. A data base was created which defined each video segment's RANGE BINS in terms of video frame numbers which were then used to generate input files for controlling the experimental presentations.

The experimenter reviewed FRACTIL video segments selected for stimuli. During the review, various squares were created with TARGA+ and overlaid on FRACTIL video segments. The squares were designed to circumscribe a specific area on the FRACTIL video segments for subject evaluation. Squares (without FRACTIL video) were recorded, each on a single laser disk frame, and the frame number was referenced to the video frame numbers of the FRACTIL video segment to which it corresponded in a separate computer file. The squares-FRACTIL video combinations were employed as subject stimuli for TSD rating responses.

During the TSD evaluation RANGE BIN, four unique squares were displayed during each evaluation, and the subjects provided a TSD rating for the imagery within the square. Whether a square partition area contained signal (a

single target) was determined randomly (probability of target present equaled 0.5), and for each condition there were 70 signal-plus-noise trials and 70 noise-only trials.

Stimulus preparation for hit-FAC response data required a different approach. The experimenter reviewed the video segments and identified a unique set of circular "truth" regions about the true targets in the scene.² A software application was created which allowed circular regions to be differentiated from all other regions of the display during subject touch responses. The circular regions were not visible in the stimuli. Subject touches within the truth regions were electronically logged as a hit, and touches not within a truth region were logged as false alarms. The resultant data were used to generate hit-FAC curves for each subject.

During each video segment presentation and prior to the presentation of the first RANGE BIN for evaluation, the subjects were presented ten seconds of imagery recorded just prior to the RANGE BIN start. This acclimation video depicted the target array area as the subject "flew" toward his targets and into the RANGE BIN for evaluation. This allowed the subjects to acclimate to the target distances and other scene conditions, and to study the imagery just as they would in an operational setting.

Because the imagery was recorded as the aircraft flew toward the target array, the ground area within the field of view narrowed and the number of targets visible in the scene decreased with decreasing range to target. This effect induced different numbers of hit opportunities for each RANGE BIN. The total numbers of hit opportunities for the Edwards imagery were 123, 152, 187, and 236 for RANGE BINS 2, 4, 6, and 8, respectively. The total numbers of hit opportunities for the Eglin imagery were 136, 191, 195, and 215, for RANGE BINS 2, 4, 6, and 8, respectively.

² True targets were identified using target array survey plots from the FRACTIL demonstration flights.

The presentation of video segments was constructed so that both hit-FAC and TSD data were collected for all four RANGE BINS. For RANGE BINS 8 and 2, some video segments were constructed to collect only one type of data. Because hit-FAC data were always collected first in each video segment and TSD data were collected for a subsequent RANGE BIN in the segment, video segments were required which facilitated only TSD data for bin 8. No hit-FAC data were required at ranges beyond bin 8. Similarly, additional video segments facilitated only hit-FAC data for bin 2. No TSD data were required for ranges closer than bin 2. When these circumstances occurred, the subjects were alerted with special visual messages.

All video segments were monochromatic and were presented on a gray-scale, monochromatic display. The order of presentation of video segments was randomized within the constraints imposed by the experimental apparatus. Four video laser disks were required to hold all of the video used for experimental presentation. The apparatus allowed the random presentation of video segments from two disks at a time. One disk contained only Edwards conditions imagery. Two disks contained only Eglin conditions imagery. One disk contained both Edwards and Eglin conditions imagery. Two stimulus sets were designed using two disks each, and each set contained approximately equal quantities of imagery from each of the two locations. The presentation order of these two stimulus sets was randomly selected, and the order of video presentations comprising each stimulus set was completely randomized.

Procedure

The general experimental technique was to present segments of operationally derived infrared video imagery to subjects who pointed out targets imbedded in the imagery, and who evaluated predefined regions of imagery for the presence or absence of targets. Each video segment consisted of ten seconds of "acclimation" video followed by two successive RANGE BIN presentations. During the first RANGE BIN, subjects provided responses which facilitated the

creation of hit-FAC curves. During the second RANGE BIN, subjects provided TSD ratings necessary for the generation of ROC curves and the calculation of d' values. The video stimuli were presented so that both types of response data were collected for four RANGE BINS and for both atmosphere-clutter conditions, resulting in eight unique experimental conditions.

Before the initiation of an experimental session, the subjects were given a complete explanation of the task. Prior to actual data collection, the subjects were shown multiple samples of FRACTIL video tape and the true targets in the imagery were pointed out by the experimenter. The subjects viewed sample video which encompassed all RANGE BINS under study. The subjects were informed that a maximum of nine targets and a minimum of one target were present in any given scene.

Following this imagery familiarization exercise, the subjects completed a practice session of procedures identical to the actual data collection trials. The subjects received feedback on performance following each practice trial. Any questions about procedures were answered and additional practice was allowed at the discretion of the subject. All subjects elected to take some additional practice.

For the data collection trials, subjects were seated a fixed distance from the video monitor which was representative of the typical viewing distance in the F-16 cockpit -- approximately 30 inches. Minor adjustments were allowed in the seating distance to accommodate various arm reach distances among the subjects. The subjects were enclosed in a semi-darkened booth throughout the experiment.

The monitor displayed instructions for the subjects to touch the screen to begin the first video segment. Upon touching the screen, the subjects viewed a segment. The segment began with ten seconds of target ingress video recorded prior to the RANGE BINS of interest for the segment. The subjects began searching for targets in the scene immediately. At the end of the ten seconds and at the start of the first RANGE BIN, a tone sounded indicating to the subjects that hit-FAC reporting should begin. No change or interruption in the video presentation occurred, however the undisplayed, predefined circular regions about

the true targets were activated at the start of the first RANGE BIN. The subject immediately began touching the screen at locations for which he believed a target to be present. He attempted to touch potential target regions in the order of his confidence that the region was a true target. That is, regions providing very strong indication of being a target were touched first, while regions which were questionable targets were touched last. The subject received aural feedback confirming the registration of each touch. These feedback tones were of shorter duration, higher pitch, and different character than the initial RANGE BIN notification tone. The subject continued to provide touch screen reports until the expiration of the RANGE BIN was indicated by presentation of the words "STOP TOUCH" in a vertical arrangement on the left and right edges of the display.

The computer and touch screen device logged all touch reports. Touches within one of the circular target regions were annotated with a number identifying the exact circular region touched, and these were considered hits. All other touch reports were considered false alarms. Multiple trial reports were sorted by their reporting order within the trial, and hit-FAC curves were plotted for each range using procedures detailed in the *Data Analysis* section of this paper. If a subject touched the same circular target area more than once, only the first occurrence was logged as a hit and the subsequent touches were ignored.

No interruption in the video segment occurred after the termination of the hit-FAC reports, and the subject continued viewing. After viewing one interim RANGE BIN (about 5 sec), four touch-screen rating choices were displayed along the left and right sides of the display, and a computer-generated square was overlaid on the video scene. The four rating choices, labeled one through four, were predefined for the subject as follows: (1) Target definitely not present, (2) target probably not present, (3) target probably present, and (4) target definitely present. Figure 14 illustrates the display.

The subjects provided a TSD rating via the touch screen for the region of the display bounded by the square. The first square was presented for two

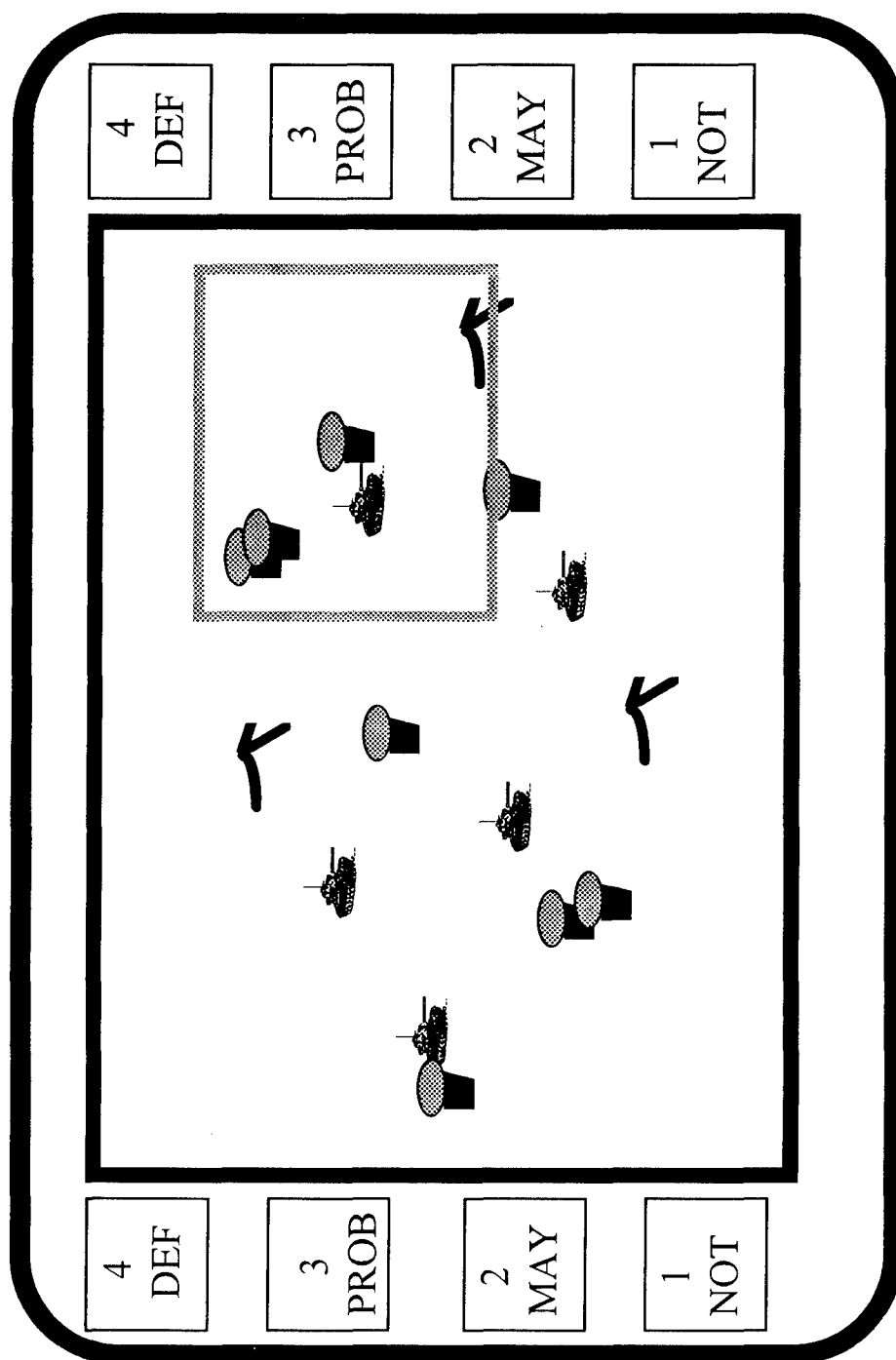


Figure 14. Illustration of typical display and menu for TSD stimuli and touch response.

sec and then replaced with a second square. The subject provided a second TSD rating for the second square, which was also displayed for two sec. Similarly, a third and fourth square were displayed for TSD reports, each for two sec. Two-sec exposure times were chosen as the minimum reasonable time in which subject reports could be logged for each polygon. Since the RANGE BIN video presentations last approximately five sec, freezing the video frame for three sec was necessary to allow reporting without exposing video of ranges closer than the RANGE BIN of interest. The video segment stopped in a freeze-frame mode at the end of the second RANGE BIN of interest until the expiration of the two-sec presentation time for the fourth square. This freeze frame time was approximately three sec in duration. Aural feedback of rating touches was again provided to the subject. The overall trial chronology is depicted in Figure 15.

After the freeze-frame video and the expiration of the final polygon display time, the video scene was blanked. The subject was presented with a message informing him to touch the screen to initiate the next video segment. Upon touching the screen, the subject was presented with another similar scenario for a new set of two RANGE BINS. This trial scenario was continued until all experimental stimuli were viewed.

Because the run time of this experiment was approximately 3.5 hours in duration, frequent breaks were allowed at the discretion of the subject. The subjects were provided with constant updates of progress toward completion via a second computer monitor. Four subjects completed both stimulus sets in the same day. Eight subjects completed one set each on two different days due to difficulties scheduling a four-hour time block. In these cases, a review of the familiarization and training procedures was conducted immediately before the start of the second stimulus set.

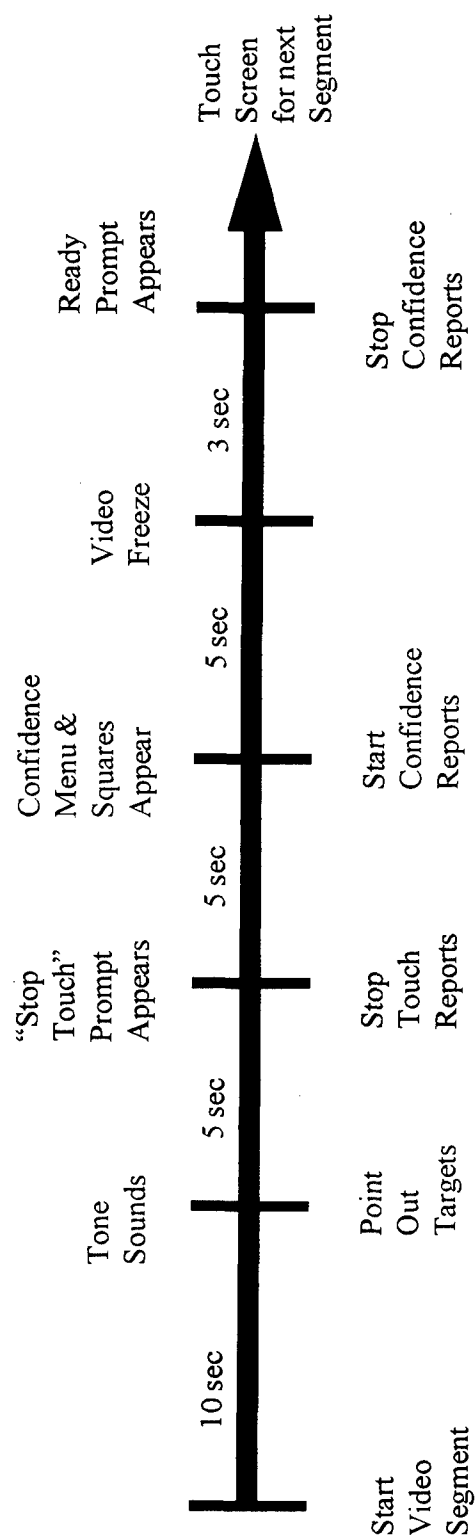


Figure 15. Stimulus presentation chronology.

Data Reduction

As indicated in the *Experimental Design* section, the dependent measures examined were (1) hit-FAC hit rate for a false alarm count of two, and (2) d' as determined from TSD procedures with the assumption of equal variance of the noise-only and signal-plus-noise distributions. The hit rate measure required only a slightly different reduction technique when applied to a human rather than an ATR. However, reduction and analysis of the TSD data revealed that the assumption of equal variance in the decision space distributions was an invalid assumption. An alternative TSD metric, d'_e , was substituted for d' to compensate for error which may be induced when using d' with unequal distribution variances. Both variables were subjected to an analysis of variance (ANOVA) after being reduced to an acceptable form.

Hit-FAC Data Reduction. The previously described technique for plotting hit-FAC curves required that confidence ratings be ranked from highest to lowest confidence to facilitate plotting the curve (See *Background* section). In order to incorporate all data for the same RANGE BIN, the ATR evaluator may merge all confidence reports from all trials and then rank the entire list. Because an ATR can resolve confidence reports to four decimal places or greater, the likelihood of two reports being equal in value and also having opposite truth characteristics is quite small. Should this event occur, the experimenter is uncertain which truth characteristic to plot first (either the hit or the false alarm) because the confidence values are equivalent. With such large numbers of reports, the difference in outcome of the plotted curve is minor.

However, since the human cannot report his confidence with the same mathematical precision and expedience as the ATR, the likelihood of equivalent ratings having opposite truth characteristics is quite high. In the hit-FAC procedure described here, the subject provided confidence ratings by the order in which potential targets were pointed out. Thus, every trial had equivalent ratings even though the confidence levels may have been very different, and there was no

way to recover these ratings as can be done with the ATR. Clearly, an alternative to the ATR data reduction procedure was necessary to generate hit-FAC curves for human subjects.

The problem was solved by employing an alternative averaging technique. Rather than ranking each individual rating, the ratings were sorted by selection order and plotted by relative fractions. All of the "first choice" ratings were examined and the quantities of hits and false alarms within the group was determined. The hit-FAC vertical axis was still incremented by the total number of hit opportunities being considered within the RANGE BIN. A point was plotted for the "first choice" group by moving up the vertical axis a number of increments equal to the hits, and then moving horizontally in the false alarm direction an amount equal to the relative portion of unity represented by the number of false alarms in the group.

For example, this experiment expected 35 "first choice" reports for each condition since there were 35 trials per condition. If for any one of the conditions 30 hits and 5 false alarms resulted, a point was plotted by counting 30 increments vertically along the hit axis, and then moving to the right $5/35$ of one increment of false alarm count.³ Following this, all of the "second choice" reports were scored and accumulated with the first choice scores.

In our example let's assume that the second choice results were 25 hits and 10 false alarms. The ordinate would be the sum of the first choice and second choice hit counts ($30 + 25 = 55$ increments), while the abscissa would be the sum of the first choice and second choice false alarms ($5/35 + 10/35 = 15/35$ or $3/7$). The cumulative processes were continued for all selection order groups and the hit-FAC curve was plotted.

Not all trials within a RANGE BIN had equivalent report-ordering quantities since subjects pointed out different numbers of targets on different trials.

³ Counting increments on the normalized hit-rate axis is equivalent to plotting a percentage of the total number of hit opportunities as calculated by: hits/hit opportunities. Since the number of false alarm opportunities may vary (decreases as choice order increases), the absolute increment count is not a convenient "short-cut" to employ for false alarm counts.

However, the relative percentages for each ordered group were still calculated and plotted. For instance, the third choice group may have had only 33 reports consisting of 20 hits and 13 false alarms. Twenty hit increments were accumulated and 13/33 (.394) false alarm units were accumulated. Figure 16 provides a graphical depiction of an example.

Employing this graphical averaging technique generated a hit-FAC plot with less resolution than that produced by the ATR, but the resolution was limited by the human reporting technique. The shape of the curve was still ascertained, and the hit rate was determined for a designated false alarm count of two.

After the hit rate was determined for each RANGE BIN for each subject, an analysis of variance (ANOVA) was conducted on the 4 x 2 repeated measures design. The analysis examined the main effects of range and of atmosphere-clutter conditions. The interaction between these two variables was also examined.

TSD Data Reduction. Mathematically determining values of d' for the TSD portion of the experiment was a simple procedure. Recall that d' is a measure of the difference between the means of the decision space distributions of noise only and signal plus noise. Also, recall that the rating experiment procedure provides values of hit rate and false alarm rate at multiple values of observer bias. A value of d' was determined at each level of bias. The number of d' estimates equals one less than the number of rating options since the total accumulated probabilities under the final rating option equals unity and represents an anchored point on the ROC curve (Green and Swets, 1966). Thus, for this experiment three d' values were determined for each RANGE BIN since a four-point rating scale was employed.

In order to easily calculate the d' values and test the assumption of equivalent variance, the hit rate and false alarm rates (the ROC curve points) were transformed into z-scores using a statistical spreadsheet function for normal

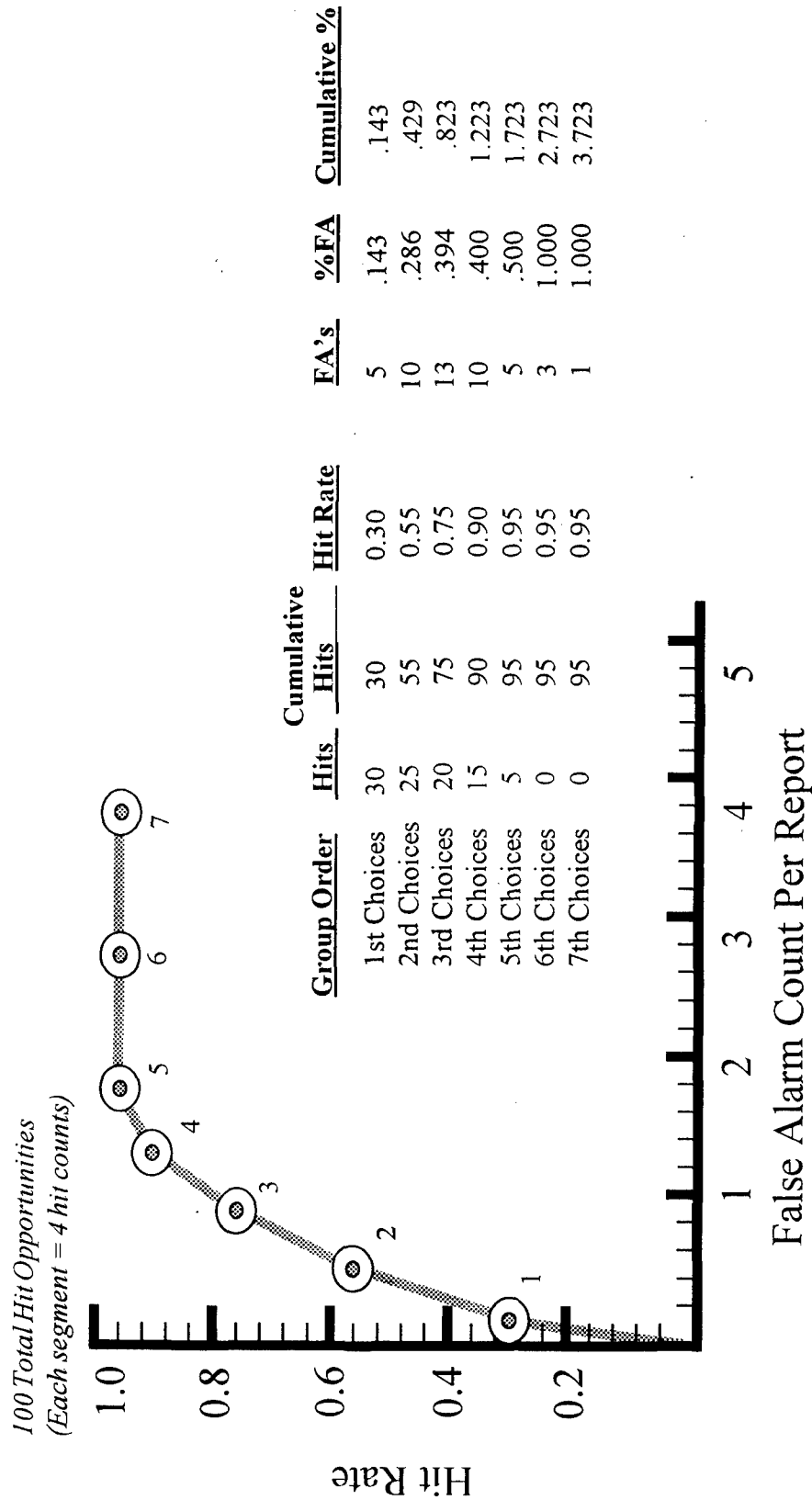


Figure 16. Example of hit-FAC curve creation from human subject data.

distributions⁴. The difference between each associated pair of false alarm rate and hit rate z-scores is d' for the bias level associated with the pairing (Macmillan and Creelman, 1991). Table 5 illustrates an example.

Table 5. Cumulative proportions of each response rating transformed into z-scores for the calculation of d' .

$z(\text{FAR}) - z(\text{hit rate}) = d'$, for each cumulated score.

	<u>Rating Response</u>			
	4	3	2	1
Hit Rate	.706	.765	.882	1.00
z-score transform	-0.541	-0.723	-1.19	n/a
False Alarm Rate	.044	.074	.162	1.00
z-score transform	1.70	1.45	0.987	n/a
d' calculation	2.24	2.17	2.17	n/a

In order to test the assumption of equivalent variance in the decision space distributions, the z-transforms were plotted on z-axes of hit rate and false alarm rate. Again employing a statistical spreadsheet function, a line for the three points was determined using the least squares method. The slope of the z-transform ROC line was noted for each subject and condition. Since over fifty percent of the lines were of slopes substantially different from one, indicating unequal variance in the distributions, d'_e was calculated as an alternative metric. The measure d'_e is less sensitive to the impact of unequal variance than d' since it gives equal weight to the units of the noise-only and signal-plus-noise distributions (Green and Swets,

⁴ The accumulated proportions for ROC curve plotting were transformed into z-scores and plotted on z-coordinates. In these coordinates the ROC curve is typically a straight line whose linearity and slope reveal the nature of the decision space distributions. Non-linearity is indicative of non-normality. An ROC slope value other than one is indicative of unequal variance in normal distributions (Macmillan and Creelman, 1991).

1966). It is described in detail for the interested reader by Green and Swets (1966).

Once d'_e was determined for all RANGE BINS for each subject, an ANOVA was conducted for the 4 x 2 repeated measures design. Finally, a correlation between the d'_e values and the hit-FAC values was determined to seek a relationship between the two.

RESULTS

Data Analysis

Subject performance was analyzed using the Statistical Analysis System (SAS), PC SAS for Windows, version 6.08, on a P5 (Pentium) personal computer. Main effects and interactions were evaluated using a significance criterion of 0.05, and simple effects F-tests were used to analyze significant interactions. Tukey's Honestly Significantly Difference Test was used to assess all pairwise comparisons.

Values of d'_e were determined from TSD ratings data. Hit-FAC hit rates were derived from each hit-FAC curve at a false alarm count of two. In a preview of the subject data, a false alarm count of two seemed to be the value near which a majority of the hit-FAC curves began to plateau. Since this region of the curve depicts the most desirable operating performance (maximum hit rate for minimum false alarms), the value of two was selected as the false alarm count from which all subject hit rates would be derived.

TSD technique (d'_e variable). The values of d'_e are greatest for those RANGE BINS representing relatively short ranges to target and the smaller d'_e values are associated with longer ranges to target. This trend is clearly depicted in Figure 17, which illustrates the interaction of RANGE BIN and ATMOSPHERE-CLUTTER for d'_e .

As the ANOVA summary in Table 6 indicates, the main effect of RANGE BIN was significant for d'_e ($p \leq .0001$). The main effect of ATMOSPHERE-CLUTTER (ATM-CLUT) was also significant ($p = .0108$), and the interaction of RANGE BIN and ATMOSPHERE-CLUTTER was significant ($p \leq .0001$).

F tests of simple effects of RANGE BIN for different levels of ATM-CLUT were conducted. There were significant differences among RANGE BIN for ATM-CLUT conditions for d'_e (Edwards, $p \leq .0001$; Eglin, $p \leq .0001$).

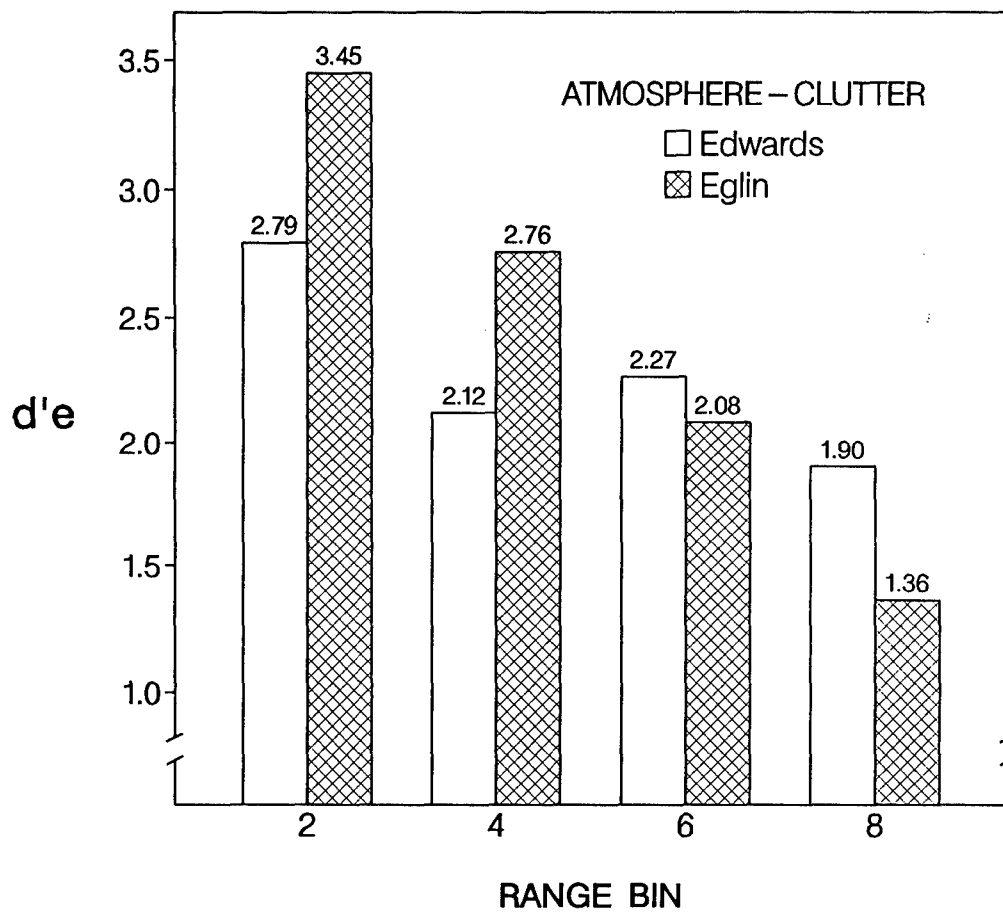


Figure 17. $d'e$ vs RANGE BIN for both levels of ATMOSPHERE - CLUTTER.

Table 6: ANOVA Summary for d'_e .

Source	df	MS	MS error	F	p
Subjects	11	0.141			
RANGE BIN	3	9.201	0.0868	105.94	0.0001
ATM-CLUT	1	0.495	0.0527	9.39	0.0108
BIN x ATM- CLUT	3	2.162	0.0474	45.61	0.0001

Pairwise comparisons of RANGE BIN means using Tukey's technique were performed for both Edwards and Eglin conditions. A summary of these results is provided in Table 7.

For the Edwards conditions, a significant difference was found between RANGE BIN 2 and all other RANGE BINS, and RANGE BIN 6 was found significantly different from RANGE BIN 8. However, the difference between RANGE BIN 4 and RANGE BIN 6 was not found to be significant, nor was the difference between RANGE BIN 4 and RANGE BIN 8 significant. For the Eglin conditions, d'_e for each RANGE BIN was significantly different from all other RANGE BINS.

A simple effects F-test of ATMOSPHERE-CLUTTER by RANGE BIN show significant differences for d'_e between the Edwards and Eglin levels for all four RANGE BINS. Table 8 summarizes these results. RANGE BIN 2, 4, and 8 were significant for ATM-CLUT ($p \leq .0001$), and RANGE BIN 6 was significant ($p = .0494$). An examination of Figure 17 reveals that RANGE BIN 6 represents an area near where the d'_e values for Edwards and Eglin "cross over".

Table 7. Results of simple effect analysis of RANGE BIN by ATMOSPHERE-CLUTTER using the $d'e$ dependent variable.

ATM-CLUT	df	MS	F	p
Edwards	3	1.726	25.71	0.0001
Eglin	3	9.637	143.58	0.0001

Pooled MSE = 0.0671 Pooled dfE = 60.8 MSD = 0.280

<u>ATM-CLUT</u>	<u>BIN</u>	<u>Mean</u>	<u>Tukey Grouping</u>
Edwards	2	2.794	A
	6	2.265	B
	4	2.122	B C
	8	1.903	C
Eglin	2	3.452	A
	4	2.760	B
	6	2.083	C
	8	1.364	D

Hit-FAC technique (HR variable). Analyses identical to those for the $d'e$ variable were conducted for the hit-FAC hit rate variable. Once again, the dependent measure is greatest for those bins representing relatively short ranges to target and the smaller HR values are associated with longer ranges to target. Figure 18 depicts the interaction of RANGE BIN and ATMOSPHERE-CLUTTER for HR and clearly reveals the trend. The ANOVA summary in Table 9 indicates that the main effect of RANGE BIN was significant for HR ($p \leq .0001$), and the main effect of ATMOSPHERE-CLUTTER (ATM-CLUT) was

Table 8. Results of simple effect analysis of ATMOSPHERE-CLUTTER by RANGE BIN using the d'_e dependent variable.

BIN	Source	df	MS	F	p
2	ATM- CLUT	1	2.593	53.21	.0001
4	ATM- CLUT	1	2.445	50.18	.0001
6	ATM- CLUT	1	0.199	4.08	.0494
8	ATM- CLUT	1	1.743	35.77	.0001

Pooled MSE = 0.0487 Pooled dFE = 43.9

significant ($p = .0260$). The interaction of RANGE BIN and ATMOSPHERE-CLUTTER was also significant ($p \leq .0001$).

The simple effects F-tests by ATM-CLUT revealed significant differences among RANGE BIN for both ATMOSPHERE-CLUTTER conditions (Edwards, $p \leq .0001$; Eglin, $p \leq .0001$). Again, pairwise comparisons of RANGE BIN means using Tukey's technique were performed and are summarized in Table 10. The results for HR were similar to those found for d'_e .

For the Edwards conditions, a significant difference was found between RANGE BIN 6 and all other RANGE BINS, and RANGE BIN 8 was also significantly different from all others. RANGE BIN 2 and RANGE BIN 4 were not found to be significantly different. As with the d'_e variable, the HR for each

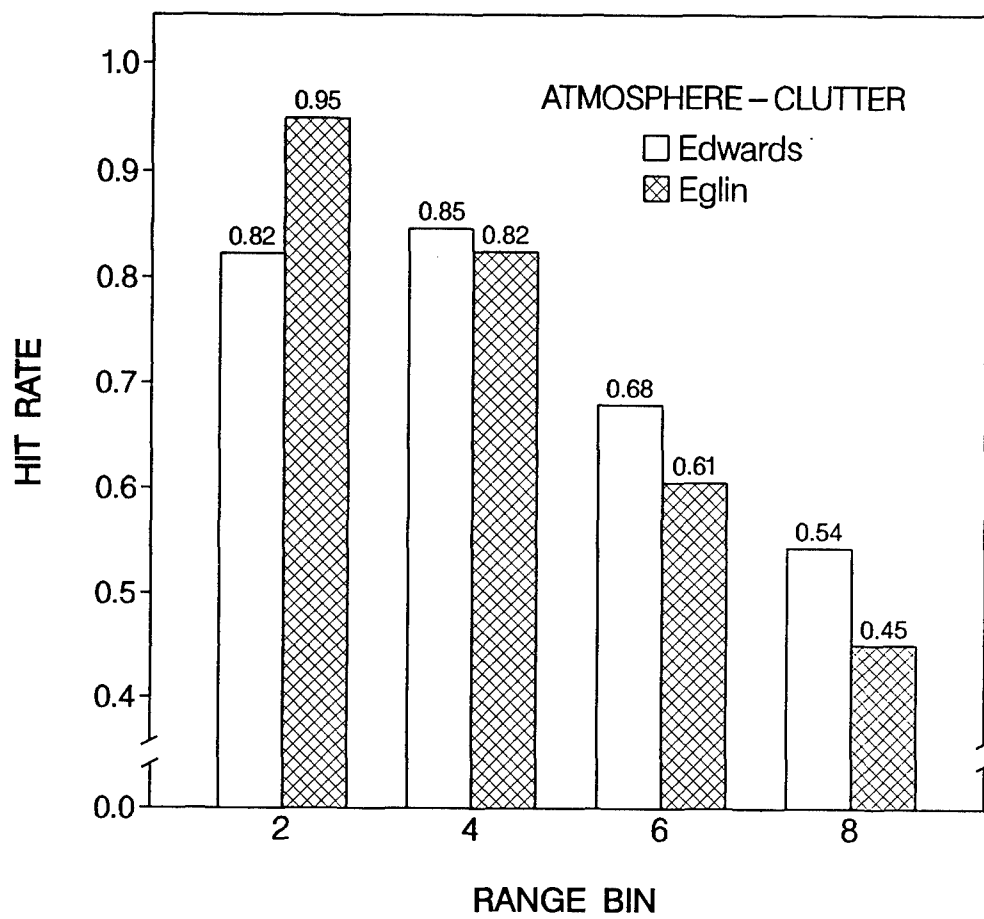


Figure 18. HIT RATE vs RANGE BIN for both levels of ATMOSPHERE - CLUTTER.

Table 9. ANOVA Summary for HR.

Source	df	MS	MS error	F	p
Subjects	11	0.0153			
RANGE BIN	3	0.7706	0.0014	557.79	0.0001
ATM-CLUT	1	0.0058	0.0009	6.61	0.0260
BIN x ATM- CLUT	3	0.0591	0.0010	59.91	0.0001

Table 10. Results of simple effect analysis of RANGE BIN by ATMOSPHERE-CLUTTER using HR dependent variable.

ATM-CLUT	df	MS	F	p
Edwards	3	0.2363	197.44	0.0001
Eglin	3	0.5933	495.70	0.0001

Pooled MSE = 0.0012 Pooled dFE = 64.3 MSD = 0.037

<u>ATM-CLUT</u>	<u>BIN</u>	<u>Mean</u>	<u>Tukey Grouping</u>
Edwards	4	0.846	A
	2	0.823	A
	6	0.680	B
	8	0.544	C
Eglin	2	0.949	A
	4	0.825	B
	6	0.606	C
	8	0.451	D

RANGE BIN under Eglin conditions was significantly different from HR for all other bins.

Simple effects F-test of ATMOSPHERE-CLUTTER by RANGE BIN show significant differences between the Edwards and Eglin levels for three of the four bins. Table 11 summarizes these results. RANGE BIN 2, 6, and 8 were significant for ATM-CLUT ($p \leq .0001$), but RANGE BIN 4 was not significant ($p = .0867$). Again, a "cross over" in the dependent measures can be observed at the insignificant comparison. RANGE BIN 4 is the cross over area for the HR variable as depicted in Figure 18.

Table 11. Results of simple effect analysis of ATMOSPHERE-CLUTTER by RANGE BIN using the HR dependent variable.

Bin	Source	df	MS	F	p
2	ATM- CLUT	1	0.0955	93.23	.0001
4	ATM- CLUT	1	0.0028	3.07	.0867
6	ATM- CLUT	1	0.0327	33.77	.0001
8	ATM- CLUT	1	0.0521	53.21	.0001

Pooled MSE = 0.0010 Pooled dFE = 43.9

Correlations. The similarity in results between the d'_e measure and the HR measure are obvious. The correlation of d'_e and HR was examined and the correlation coefficient (r) was determined based on the Pearson product-moment

correlation. Correlations were examined separately for each of the ATMOSPHERE-CLUTTER conditions and for the combination of the two.

A correlation coefficient of $r = 0.89$ ($R^2 = 0.79$) was determined overall for the combined Edwards and Eglin results, indicating a moderate positive correlation. Figure 19 depicts the regression analyses for combined ATMOSPHERE-CLUTTER levels. A very high positive correlation was found for the Eglin conditions ($r = 0.99$, $R^2 = 0.98$), but only a low moderate correlation was found for the Edwards conditions ($r = 0.62$, $R^2 = 0.39$). The regression analyses for Edwards and Eglin are presented in Figure 20 and Figure 21, respectively.

Only the Edwards regression line failed the test for significance ($p = 0.3735$), and Figure 20 reveals a single point (Edwards RANGE BIN 4) which largely accounts for the failure. The same single point is observed in the combined analysis (Figure 19) as not conforming to the well defined linear arrangement of the other points. Figure 22 depicts a combined regression analysis omitting the point associated with Edwards RANGE BIN 4, and the associated correlation coefficient is $r = 0.99$ ($R^2 = 0.98$).

The dependent measures were assumed to vary linearly with RANGE BIN, and regression analyses were conducted to examine RANGE BIN on d'_e and HR. Figure 23 depicts the regression analysis of RANGE BIN on d'_e for Edwards and Eglin. For Edwards, $R^2 = 0.74$, indicating that 74 % of the variance in d'_e can be accounted for by RANGE BIN. For Eglin, $R^2 = 0.99$, indicating that virtually all of the variance of d'_e for Eglin can be accounted for by RANGE BIN. The slopes of the d'_e regression lines are -0.13 for Edwards and -0.35 for Edwards. Figure 24 depicts the regression analysis of RANGE BIN on HR for Edwards and Eglin. For Edwards, $R^2 = 0.85$, and for Eglin, $R^2 = 0.99$. The slopes of the HR regression lines are -0.050 for Edwards and -0.086 for Eglin.

In each case the slope of the Edwards line is shallower than that of Eglin, with the cross-over point falling between RANGE BIN 4 and RANGE BIN 6. It

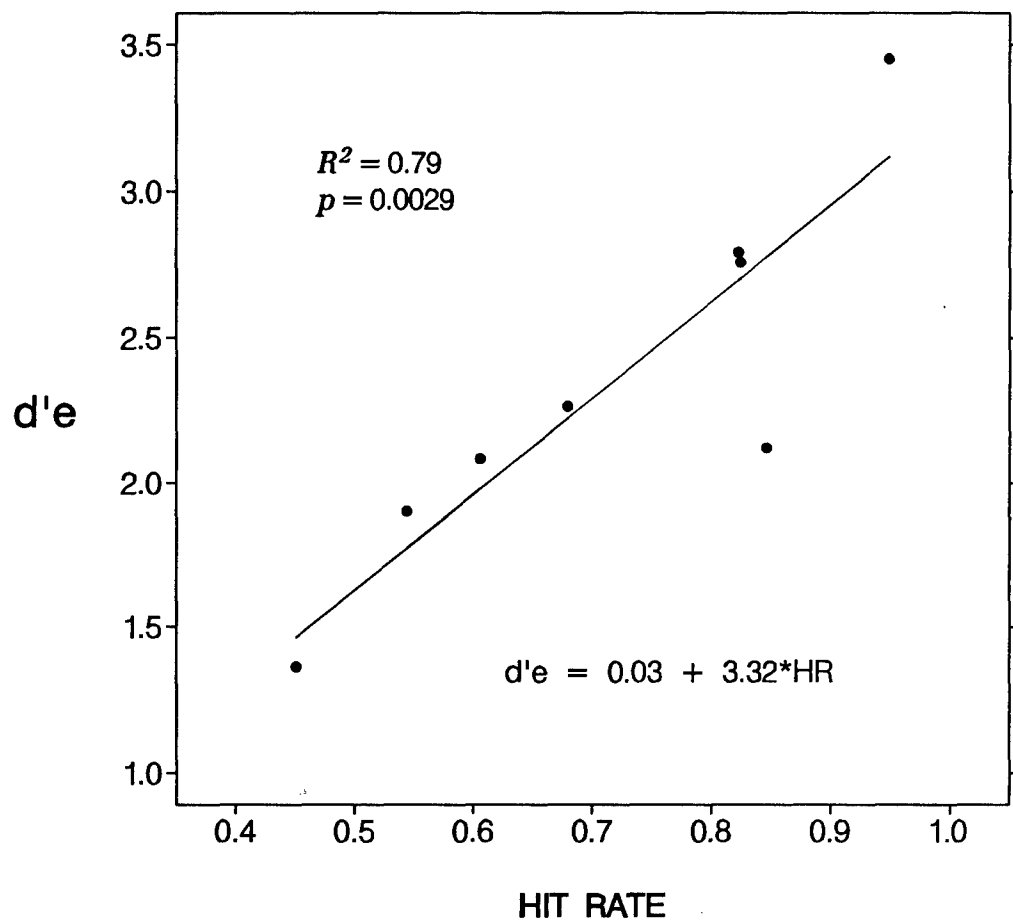


Figure 19. Linear Regression of d'e on HIT RATE for all points.

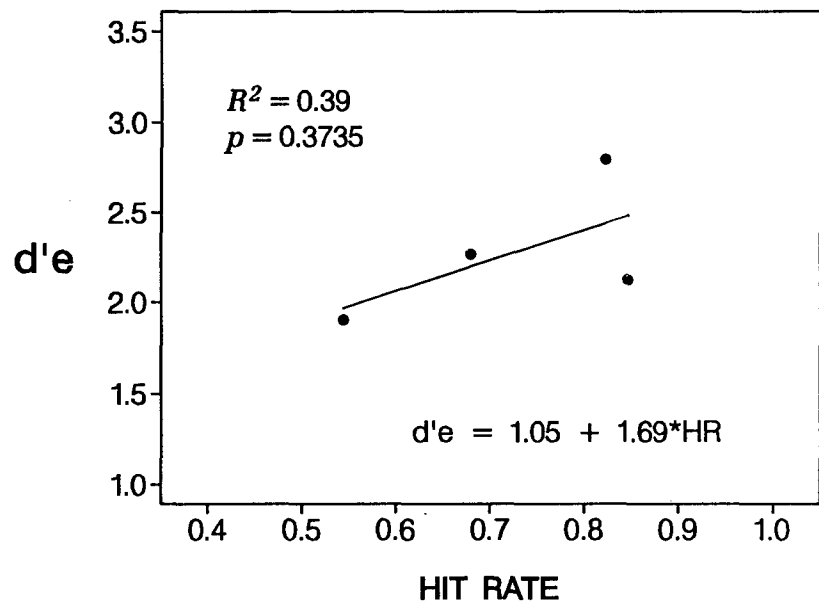


Figure 20. Linear Regression of $d'e$ on HIT RATE for Edwards.

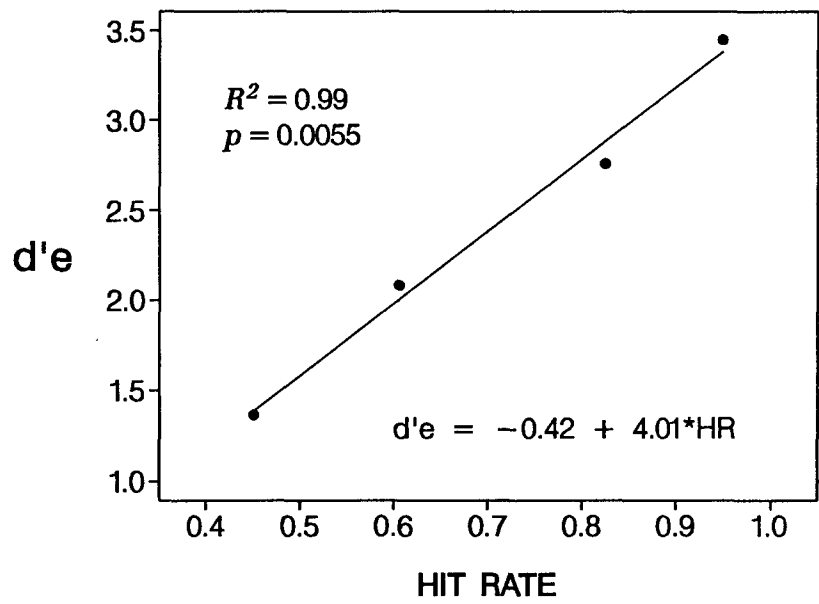


Figure 21. Linear Regression of $d'e$ on HIT RATE for Eglin.

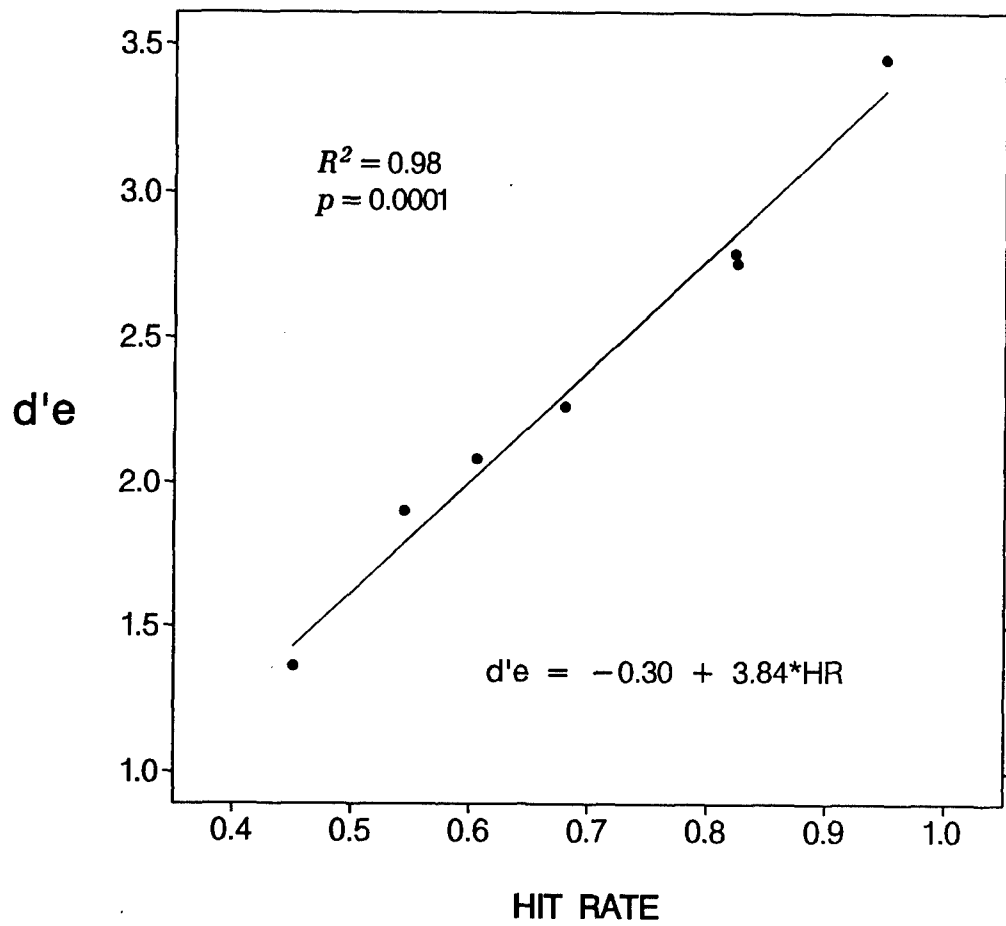


Figure 22. Linear Regression of $d'e$ on HIT RATE for all points except Edwards BIN=4.

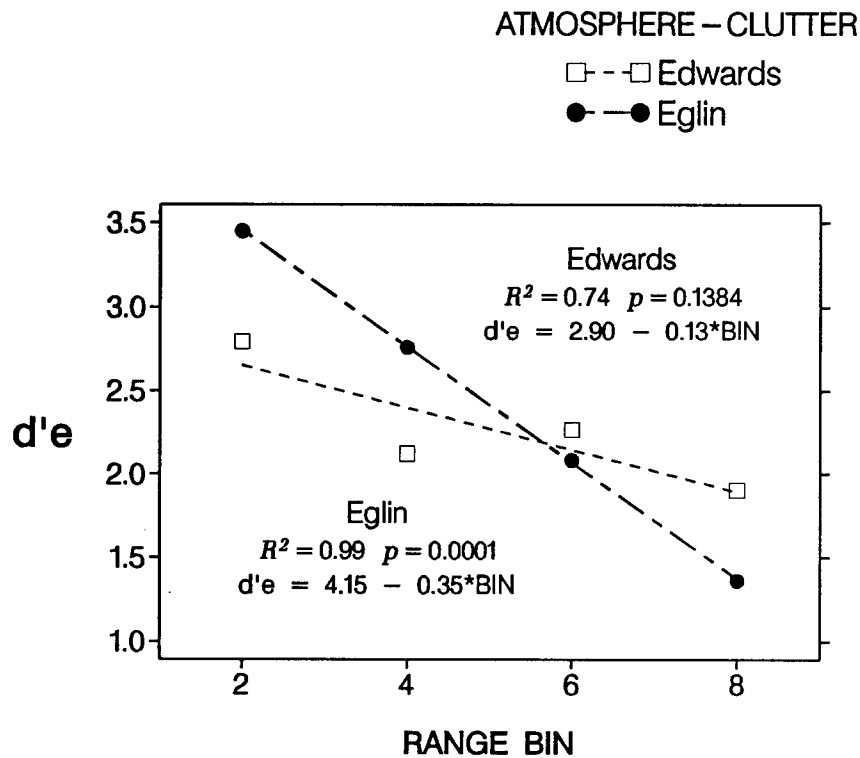


Figure 23. Linear Regression of d'e on RANGE BIN for both levels of ATMOSPHERE – CLUTTER.

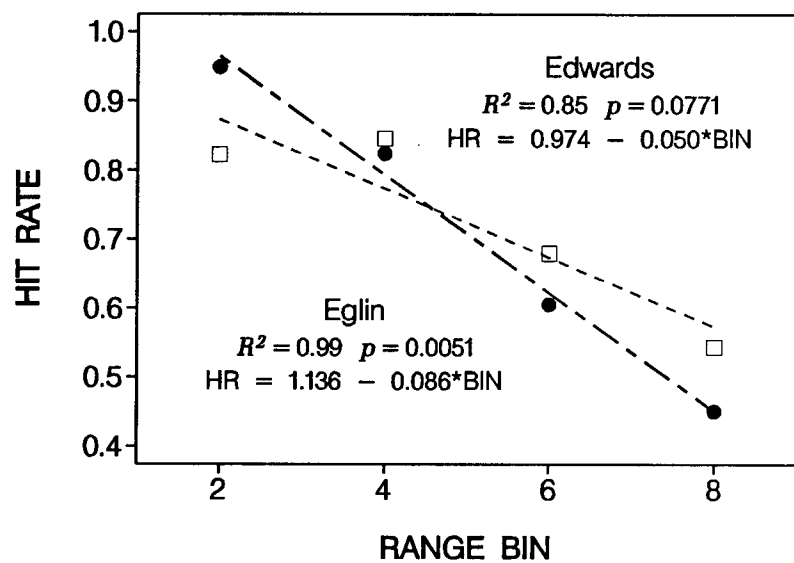


Figure 24. Linear Regression of HIT RATE on RANGE BIN for both levels of ATMOSPHERE – CLUTTER.

is interesting to note that, with the exception of RANGE BIN 4, the relative positions of the data points for each RANGE BIN are remarkably similar between the two dependent measures, and the relative magnitudes of the distances between Edwards and Eglin data points for each RANGE BIN are comparable. In each case, the Eglin data indicates strict dependence of the performance measures upon RANGE BIN, while the Edwards data is less robust. The failure of the Edwards data to equal the linearity of the Eglin data seems to be largely due to the non-conforming data point of RANGE BIN 4.

DISCUSSION

The experimental results confirmed three of the experimental hypotheses and exposed some unexpected trends. For convenience, each hypothesis is restated and followed by a discussion of the relevant experimental evidence. Additional discussion follows the review of hypotheses.

Hypotheses

Hypothesis: *Performance will degrade significantly with increased range to target under both ATMOSPHERE-CLUTTER conditions.* This hypothesis was confirmed. Examination of the bar charts of Figures 17 and 18 and the simple effects results summarized in Tables 6 and 9 reveal the effect.

For the Eglin condition, both d'_e and HR steadily decrease with increased RANGE BIN. Each bin, for each of the dependent variables, is in a separate Tukey grouping indicating a significant difference among all bins. The bar graphs suggest a strong linear relationship between range to target and both dependent variables.

The Edwards results also support the hypothesis but are less robust in depicting consistent linear relationships than are the Eglin results. For the d'_e measure, the highest value for Edwards conditions is associated with RANGE BIN 2 (2.79), and the lowest value is associated with RANGE BIN 8 (1.90). However, no significant difference is found with d'_e between RANGE BIN 4 and RANGE BIN 6 (2.12 and 2.27, respectively), nor between RANGE BIN 4 and RANGE BIN 8 (2.12 and 1.90, respectively). Examining the HR measure for Edwards conditions reveals a similar trend of decreasing HR value with increasing range. Significant decreases in HR are noted for RANGE BIN 4, 6, and 8, but no significant difference is found between RANGE BIN 2 (0.82) and RANGE BIN 4 (0.85).

The obvious disparity between the d'_e measures and the HR measures for the Edwards conditions centers about the RANGE BIN 4 results. As noted in the *Results* section, the dependent measures for this bin generated a point in the regression analysis which deviated markedly from the linear arrangement of all other data points. This suggests that either the d'_e value or the HR value for Edwards RANGE BIN 4 is anomalous. Stimuli, raw data, and truth tables for this bin were reviewed thoroughly for accuracy and consistency. No errors nor inconsistencies in stimuli or procedures were noted.

If the Edwards RANGE BIN 4 results are not an anomaly, it must be assumed that some combination of factors contributed to decreased performance in that bin alone under the TSD data collection technique, or that some combination of factors resulted in enhanced performance for RANGE BIN 4 alone under the hit-FAC technique, or that some combination of these two events occurred. Inconsistent selection of TSD stimuli could lead to a performance decrement if the selected square partitioned areas for Edwards RANGE BIN 4 were significantly more difficult to judge than all other selected areas. Subjective assessment after the experiment revealed no obvious differences in the stimuli for Edwards RANGE BIN 4 in comparison with other conditions. Similarly, enhanced performance with the hit-FAC technique could result if easily judged image segments were selected for that condition alone. However, since the same video segments were used for the other bins as well, this possibility is dismissed.

Examination of individual subject data reveals that seven of the twelve subjects provided responses resulting in Edwards RANGE BIN 4 d'_e values less than or equal to RANGE BIN 6 values. Three subjects were substantially lower for RANGE BIN 4 than for RANGE BIN 6. These three included one of the novice subjects and one subject who was suffering from a common head cold and who's overall performance was significantly poorer than the average for the group. Evaluating the data with these two individuals removed still left five of ten subjects performing poorer for Edwards RANGE BIN 4 than for RANGE BIN 6 and did

not make the difference between the two RANGE BINS significant. Further, the impact of the subject cull on Edwards HR was not remarkable.

The cause of the disparity remains unresolved. No definitive judgment can be made regarding the results for RANGE BIN 4, and further research may be required to resolve the disparity. However, the general trend of decreased performance with increased range to target is supported. This general result is not surprising in light of previous target detection research relating target detection performance to target visual angle.

Hypothesis: *The performance degradation with increased range under the Eglin conditions was expected to be more severe than that under Edwards conditions.* This hypothesis was confirmed. The trend is obvious from the regression analyses of Figures 23 and 24. Examination of the regression lines reveals a steeper regression line slope for Eglin with both d'_e and HR. The trend can also be observed in the bar graphs of Figures 17 and 18.

The deleterious effect of humidity on infrared energy sensing is well established and is easily observed in the Eglin stimuli. Water in the atmosphere absorbs energy in the infrared band, and greater ranges to target result in greater absorption of IR energy from the observed target area. The extremely high levels of humidity present during the collection of the Eglin imagery is the most likely cause of the steep drop in performance under Eglin conditions. The Edwards imagery was collected in very dry conditions and suffered very little degradation from IR energy absorption. A "flatter" performance across ranges is the result. However, atmospheric conditions including humidity could not be separated from other conditions such as visual and thermal clutter, and the results strictly support the impact of the atmosphere-clutter combination variables.

Hypothesis: *At closer ranges, detection performance was expected to be nearly equivalent for the two ATMOSPHERE-CLUTTER conditions.* This hypothesis was not supported. At closer ranges, detection performance under the

Eglin conditions was superior to that under Edwards conditions. Again, the trend can be observed in the bar graphs.

In RANGE BIN 2, significant differences resulted between the ATMOSPHERE-CLUTTER levels for both d'_e and HR. The Eglin condition produced higher values of each dependent measure. These results were unexpected. The impact of atmospheric humidity and ground clutter were expected to be small at close range where target visual angles are relatively large. If the impact of humidity is considered to be minor at the close ranges, the thermal and visual clutter typical of the Edwards condition is the most likely cause of the poorer performance under those conditions. The Eglin imagery distinctly lacked significant thermal clutter.

Hypothesis: *A linear relationship exists between the d'_e variable and the hit-FAC hit rate variable.* This hypothesis was confirmed. The relationships are depicted graphically in Figures 19, 20, 21, and 22.

As pointed out in the Results section, a strong linear correlation was found between the d'_e dependent variable and the HR dependent variable. The relationship is remarkably robust when the suspected anomaly of Edwards RANGE BIN 4 is omitted, as Figure 22 depicts.

The linear relationship is also very robust for the Eglin conditions (Figure 21), but substantially less so for Edwards (Figure 20). Again, the suspected anomaly degrades the linear relationship for the Edwards data, but the linear nature of the remaining points is obvious from examination of Figure 20.

The equation for each regression line is presented as a potential model for the task defined in this experiment. Estimates of d'_e may be obtained from a specified hit-FAC hit rate when performance is derived from the same stimulus set.

Summary

Using a different plotting technique, as in Figures 25 and 26, makes interpretation of the interaction between RANGE BIN and ATMOSPHERE-

CLUTTER easier.⁵ Standard error bars have been added to depict the variability of the means. Viewing these two figures, the steeper performance drop for Eglin is obvious for both dependent measures. It is also easier to note the deviation from linearity of the Edwards results.

Figures 23 and 24 depict the regression analyses for RANGE BIN on the dependent variables. The only notable deviation is for Edwards RANGE BIN 4. With the $d'e$ variable, the point representing Edwards RANGE BIN 4 lies below the regression line and causes the line to be "pulled" away from the linearity of the other three Edwards points. Conversely, when considering the HR data points, Edwards RANGE BIN 4 lies well above the regression line and seems to shift the regression line upward and away from the linearity of the remaining three Edwards HR points.

Alternatively, linearity is also observed for the Edwards HR variable for RANGE BIN 4, 6, and 8, and the RANGE BIN 2 point could be considered an anomaly. This scenario could be explained by the combination of thermal clutter effects and reduced opportunities for target hits in the Edwards RANGE BIN 2. That is, as the field of view of the target array area narrowed with shorter range to target, fewer targets were visible in the scene and, thus, fewer hit opportunities were available. If subjects correctly pointed out all available targets in the scene and still had time remaining to search more diligently in the scene clutter, higher numbers of false alarms might result from the subjects' desire to successfully identify all targets. Additionally, depending upon the random presentation of video segments, some subjects may have developed an expectation of high numbers of target opportunities which were characteristic of longer ranges but not the closest ranges. This effect would be present in the Edwards data more than the Eglin data due to the high level of thermal clutter at Edwards.

⁵ The dashed lines used to connect the points do not imply a known continuous function. Although some function must exist for performance along the axis representing range to target, only the discrete points plotted for each of the four range bins are known.

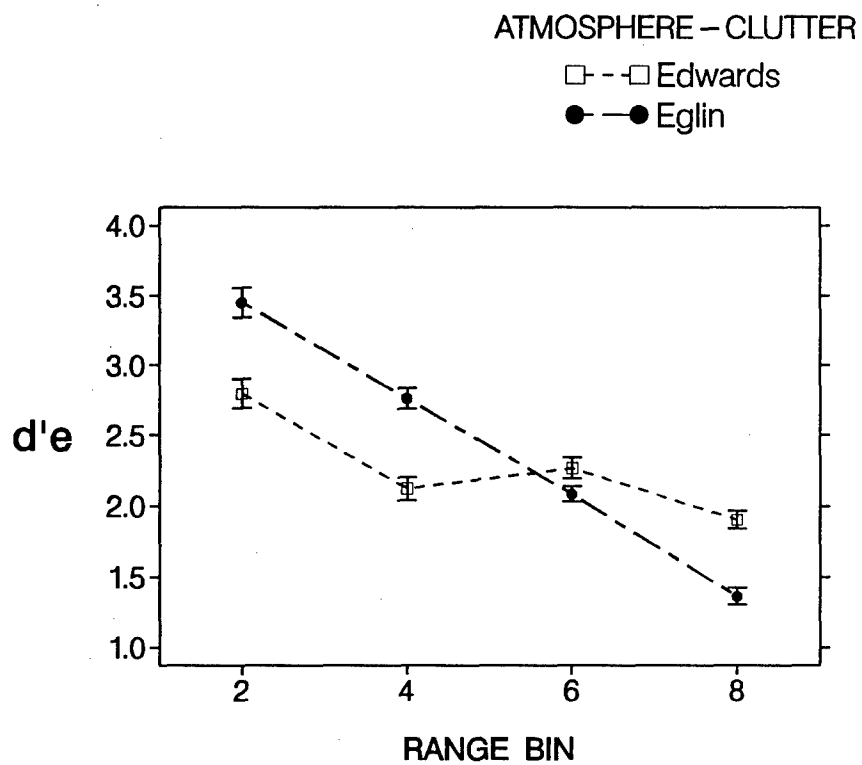


Figure 25. d'e as a function of RANGE BIN and ATMOSPHERE - CLUTTER with standard error of mean.

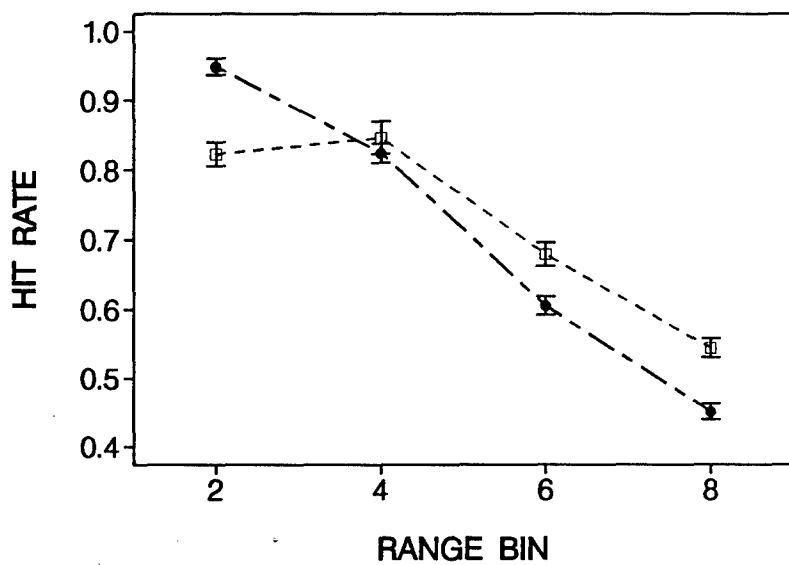


Figure 26. HIT RATE as a function of RANGE BIN and ATMOSPHERE - CLUTTER with standard error of mean.

However, subjects were required to preview video segments encompassing all bins, and each was briefed that at closer ranges the opportunities for target hits would be fewer than at longer ranges. Further, the previously described scenario would dictate a sudden and dramatic rise in false alarm occurrences at some value of confidence ordering in the hit-FAC data. A review of the raw data reveals no remarkable trend supporting this theory. These facts, along with the deviant nature of the Edwards RANGE BIN 4 point in all regression analyses, highlight the Edwards RANGE BIN 4 point as the anomaly rather than Edwards RANGE BIN 2. Future employment of the hit-FAC technique should include briefing procedures to avoid this potential problem of variable target hit opportunity.

Other potential problems for the employment of the TSD technique and the hit-FAC technique concern the size of the TSD partitioned areas (TSD squares) within the imagery and the size of the "hit" region around the true targets in the hit-FAC procedure. In this experiment, the square TSD partitions overlaid on the imagery were roughly maintained at 25 times the area consumed by a typical, single target in the scene. The circular hit regions around the true targets for the hit-FAC procedure were roughly maintained at 9 times the typical target area. Due to the manual procedures required to generate these regions, precise control of area coverage was not possible. Due to occasional movement of the overall image and a hit region update rate of 1 Hertz, a few circular hit regions were enlarged slightly to ensure coverage of true targets throughout the RANGE BIN presentation. Suggested follow-on research might investigate the impact of varying the TSD partitioned area sizes and the hit-FAC hit region sizes.

CONCLUSIONS

This experiment was highly successful. The objectives were met. A good baseline of human performance of the prescribed target detection task was established with two different techniques and two different dependent measures, d'_e and hit-FAC hit rate. The similarity in performance trends and the correlations of the two dependent variables strongly suggest a linear relationship between d'_e and hit-FAC hit rate.

A strong interaction between range to target (RANGE BIN) and ATMOSPHERE-CLUTTER conditions was verified. Three of the expected performance trends were verified and one was not supported. Target detection performance changed as a function of range to target and as a function of ATMOSPHERE-CLUTTER conditions. Generally, performance decreased with increased range to target, and performance degraded more severely in the Eglin ATMOSPHERE-CLUTTER conditions (very high humidity, low thermal clutter). However, detection performance under these conditions was superior to the performance under Edwards conditions (very low humidity, high thermal clutter) at close range to target (2 to 3 km).

The hit-FAC technique was shown to be a viable technique for evaluating human performance of a target detection task in a manner similar to that used to evaluate electronic target detection devices. The results of human performance as described by the hit-FAC hit rate statistics will be employed in follow-on research to compare the target detection performance of automatic target recognition devices with the human performance baseline. These comparisons will facilitate logical decisions about the current state of automatic target recognition technology and the acquisition of that technology for military employment.

Application of the hit-FAC technique to additional human target detection evaluations is now conceivable. Operational imagery which was previously of

limited utility in conducting laboratory target detection evaluations can now be used for such studies. The problems associated with the presence of multiple targets imbedded within a dynamic scene can be circumvented. Most promising is the relationship discovered between the hit-FAC hit rate and d'_e . Estimates of d'_e may be derived from hit-FAC data based on more elaborate models which may be derived in follow on research, and d'_e measures allow comparisons in performance independent of observer bias.

Certainly, further experimentation is necessary to establish a more general model which may apply in a universal manner to variable stimulus sets. It may be possible to derive a family of linear relationships between d'_e and HR based upon stimulus imagery characteristics such as target contrast, clutter, and other quality metrics. However, some characteristics such as clutter are difficult to describe in a deterministic manner. Additional methods for quantifying visual characteristics may be required before the proposed family of relationships can be established.

The prospects for advancing the techniques of evaluating human performance of target detection tasks are very good. This research has presented a method of bridging the gap between different evaluation techniques, and it opens numerous opportunities for further research efforts. As the interface between human and machine blurs, common evaluation metrics will be required to determine the most efficient methods of operating. This research represents one small step in that direction.

APPENDIX
SAMPLE ROC AND HIT-FAC CURVES

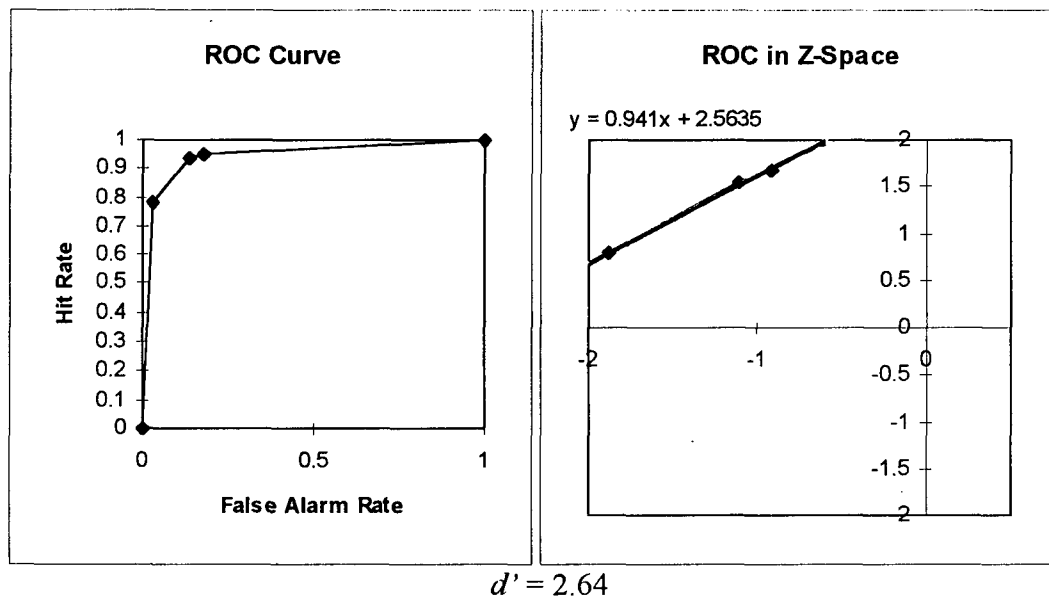


Figure 27. Sample ROC curves for Edwards RANGE BIN 2.

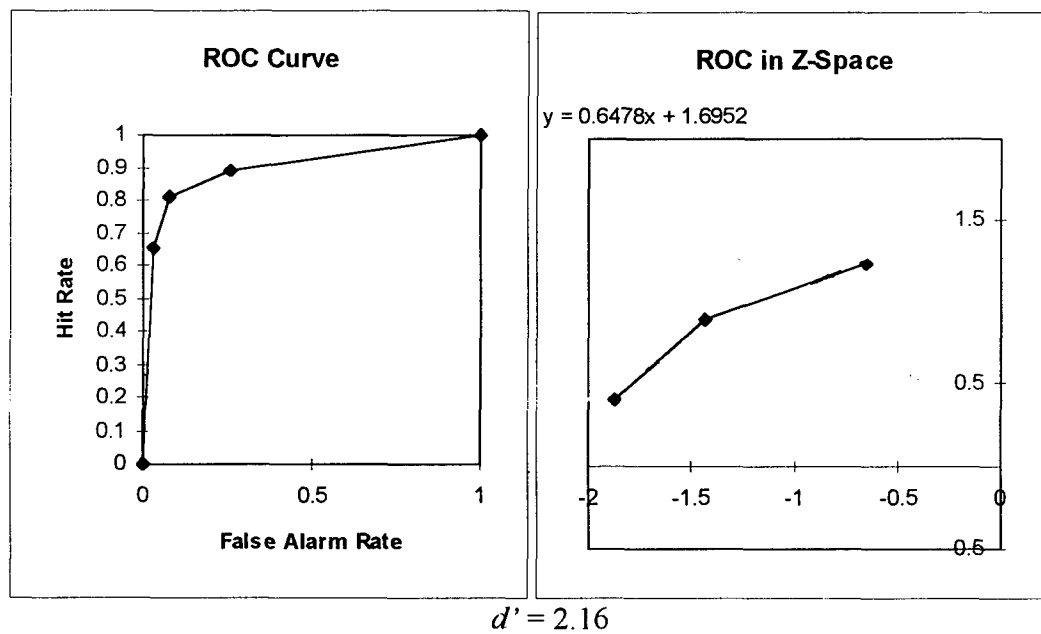


Figure 28. Sample ROC curves for Edwards RANGE BIN 4.

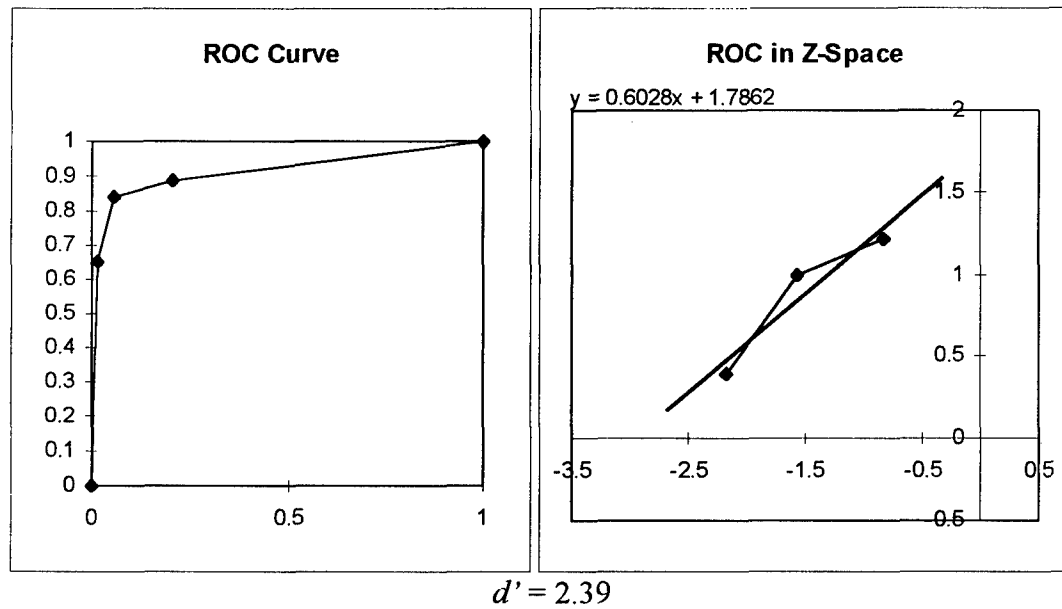


Figure 29. Sample ROC curves for Edwards RANGE BIN 6.

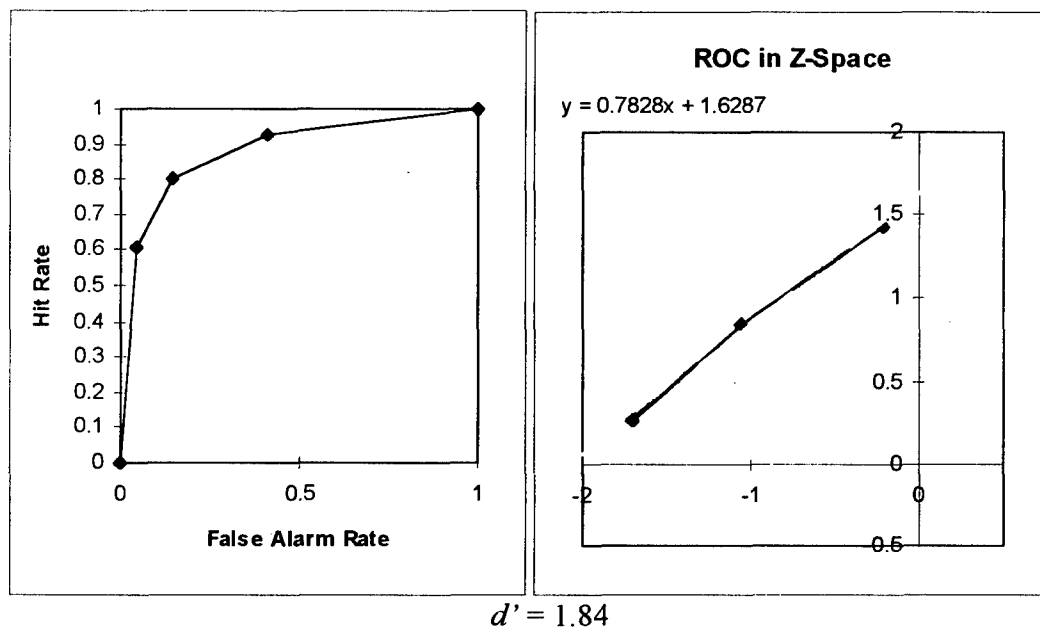


Figure 30. Sample ROC curves for Edwards RANGE BIN 8.

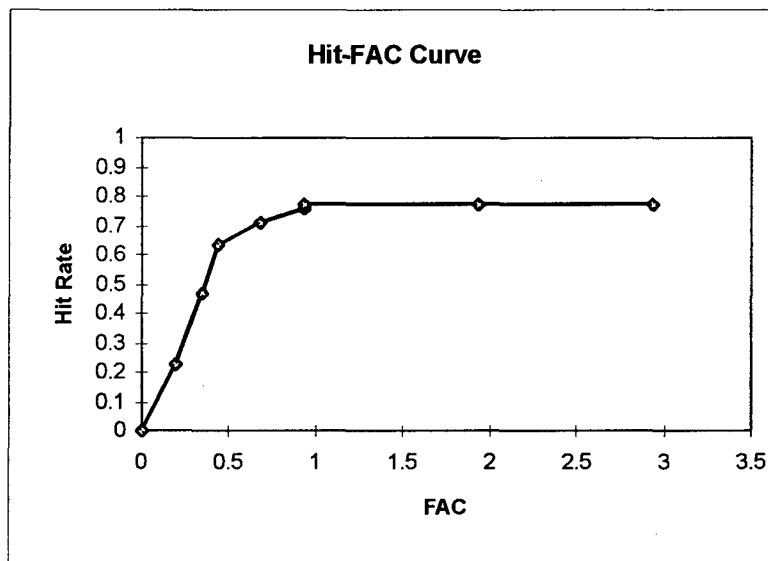


Figure 31. Sample hit-FAC curve for Edwards RANGE BIN 2.

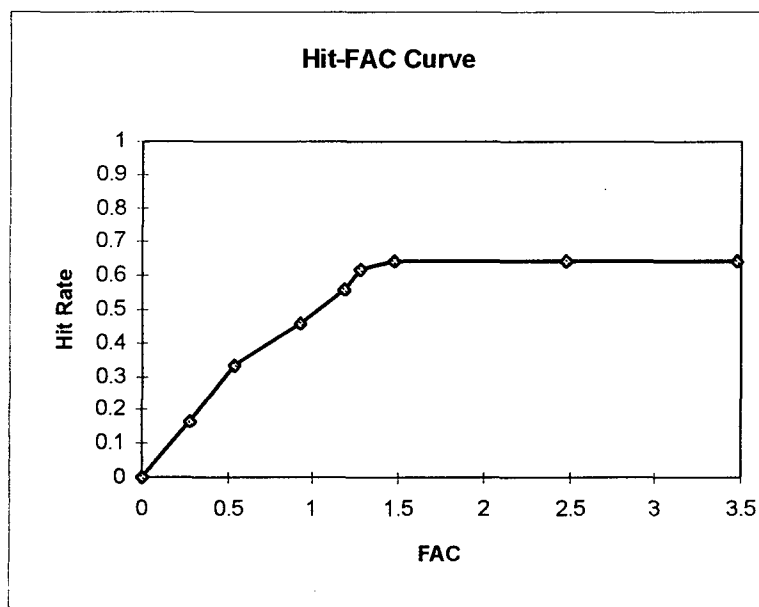


Figure 32. Sample hit-FAC curve for Edwards RANGE BIN 4.

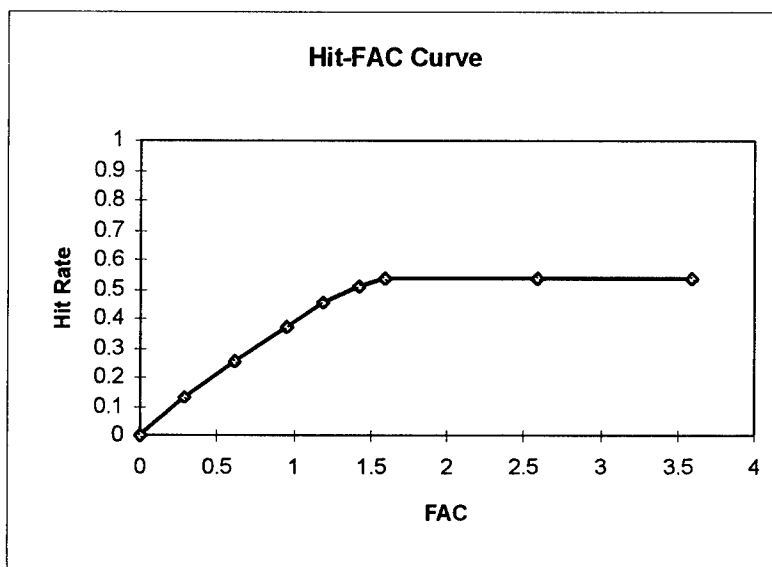


Figure 33. Sample hit-FAC curve for Edwards RANGE BIN 6.

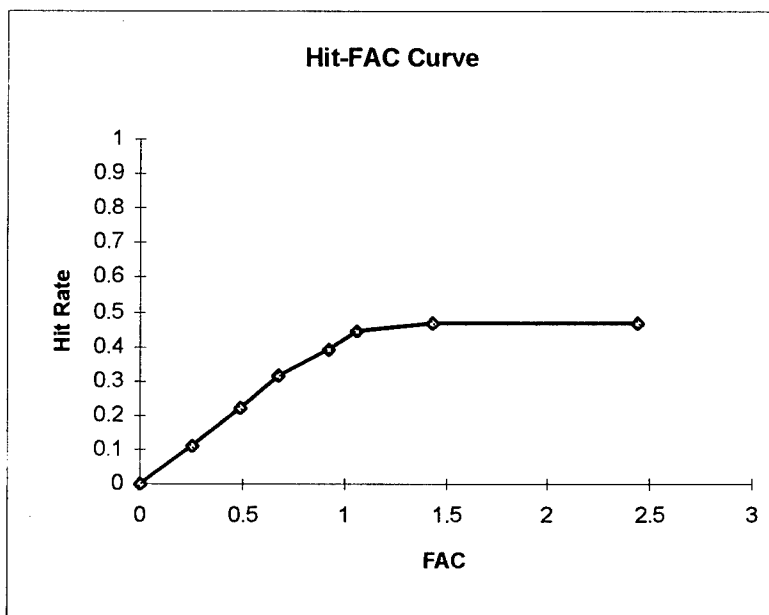


Figure 34. Sample hit-FAC curve for Edwards RANGE BIN 8.

REFERENCES

- Astley, S.M., Taylor, C.J., Boggis, C.R.M., Asbury, D.L., and Wilson, M., 1993, Cue Generation and Combination for Mammographic Screening, in Brogan, D., Gale, A., and Carr, K., (Eds), *Visual Search 2*, Burgess Science Press, Basingstoke, UK.
- Automatic Target Recognizer Working Group (ATRWG), 1986, Target Recognizer Definitions and Performance Measures, *ATRWG 86-001*, U.S. Army Missile Command, Redstone Arsenal, AL.
- Barnes, M.J., 1978, Display Size and Target Acquisition Performance, Defense Technical Information Center, Alexandria, VA.
- Beck, J., 1993, The British Aerospace Lecture: Visual processing in texture segregation, in Brogan, D., Gale, A., and Carr, K., (Eds), *Visual Search 2*, Burgess Science Press, Basingstoke, UK.
- Clark, L., 1994, Personal communication, October 22, 1994, on subject of automatic target recognition, laboratory testing of ATR's, and the generation of hit-false alarm count curves.
- Clark, L., and Westercamp, L., Personal communication, December 5, 1994, on subject of types of ATR's and experimental testing procedures.
- Egan, J.P., 1975, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, NY.
- Egan, J.P., Schulman, A.I., and Greenberg, G.Z., 1964, Operating Characteristics Determined by Binary Decisions and by Ratings, in Swets, J.A., (Ed), *Signal Detection and Recognition by Human Observers*, John Wiley & Sons, Inc., New York, NY.
- Gescheider, G.A., 1976, *Psychophysics Method and Theory*, John Wiley & Sons, Inc., New York, NY.
- Green, D.M., and Swets, J.A., 1966, *Signal Detection Theory and Psychophysics*, John Wiley & Sons, Inc., New York, NY.
- Macmillan, N.A., and Creelman, C.D., 1991, *Detection Theory: A User's Guide*, Cambridge University Press, New York, NY.
- Overington, I., 1976, *Vision and Acquisition*, Pentech Press, London, UK.

Ozkaptan, H., 1979, Evaluation of the Utility of the Theory of Signal Detection for Target Acquisition Studies, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Swets, J.A., 1964, *Signal Detection and Recognition by Human Observers: Contemporary Readings*, John Wiley & Sons, Inc., New York, NY.

Turner, S.L., Purvis, B.D., O'Hair, M., Malek, D., and Reynolds, M., in press, Performance and Workload Assessment of F-16 LANTIRN Automatic Target Recognizer System, Defense Technical Information Center, Alexandria, VA.

Wilson, D.L., 1992, Theory of Signal Detection and Its Application to Visual Target Acquisition: A Review of the Literature, Defense Technical Information Center, Alexandria, VA.

Zelnio, E.G., 1987, ATR Paradigm Comparison with Emphasis on Model-based Vision, Wright Laboratory Target Recognition Branch, Wright-Patterson Air Force Base, OH.